LUNDS UNIVERSITET
Ekonomihögskolan

Institutionen för informatik

Hantering av utmaningar med datakvalitet inom Big Data Analytics

En kvalitativ studie som beskriver hur organisationer hanterar utmaningar med datakvalitet inom Big Data Analytics

Kandidatuppsats 15 hp, kurs SYSK16 i Informatik

Författare: Frida Carlsten
Sofia Nyberg
Emelie Uddenäs

Handledare: Björn Svensson

Rättande lärare: Markus Lahtinen, Paul Pierce

ENGELSK TITEL: Managing data quality challenges within Big Data Analytics: A qualitative study describing how organisations are managing challenges within data quality in Big Data Analytics

FÖRFATTARE: Frida Carlsten, Sofia Nyberg och Emelie Uddenäs

UTGIVARE: Institutionen för informatik, Ekonomihögskolan, Lunds universitet

EXAMINATOR: Christina Keller, Professor

FRAMLAGD: maj, 2020

DOKUMENTTYP: Kandidatuppsats

ANTAL SIDOR: 82

NYCKELORD: Big Data Analytics, Big Data, Datakvalitet, Data Analytics

SAMMANFATTNING (MAX. 200 ORD):
Big Data Analytics är processen som används vid analyser av stora mängder data i syfte att upptäcka samband och mönster av värde som kan ligga till grund för framtida beslutsfattanden. En av de främsta utmaningarna som omnämns i litteraturen är att säkerställa kvaliteten på den data som används vid analyser. Syftet med denna uppsats är att beskriva organisationers hantering av utmaningar med datakvalitet inom Big Data Analytics. De utmaningar vi identifierat i litteraturen är: Information Completeness, Data Accuracy, Data Currency, Data Deduplicaton, Data Consistency, urval av Data Samples, felmarginaler, datarengöring samt ledningens stöd och engagemang. För att kunna besvara forskningsfrågan genomfördes en kvalitativ studie bestående av fem intervjuer med totalt sju respondenter. Resultatet visar på att det inte finns någon enhetlig hantering av respektive utmaning som alla organisationer tillämpar. Detta kan grunda sig i att respondenterna inte angav samma anledningar till utmaningarnas uppkomst, vilket skulle kunna bero på att de är verksamma inom skilda branscher och har olika yrkesroller.

Innehåll

Tabeller

1 Introduktion

Detta inledande kapitel har som syfte att presentera en inledande bakgrund och beskriva uppsatsens problemområde. Därefter redogörs uppsatsen syfte, forskningsfråga och avgränsningar.

1.1 Bakgrund

Det genereras dagligen enorma mängder av data och denna mängd fortsätter öka för varje dag som går (Gaikwad, Nale & Bachate, 2016). Kvaliteten på data har fått en vital roll i både beslutsfattande och operativa processer inom organisationer (Batini & Scannapieca, 2006). Dålig kvalitet på data är kostsamt för organisationer och statistik visar att dålig datakvalitet kostar amerikanska företag 600 miljarder dollar per år eller 20-35% av deras intäkter (Fan, 2015). Trots det så är problemet med dålig datakvalitet inom Big Data Analytics historiskt inte löst (Fan, 2015). Vidare är det enligt Gaikwad, Nale och Bachate (2016) nödvändigt att studera de tekniker som används för att analysera denna stora mängd data för att kunna hantera de utmaningar som Big Data Analytics medför.

Big Data Analytics gör det möjligt att undersöka och analysera enorma mängder data för att identifiera korrelationer, mönster och andra insikter (SAS, n.d.). Begreppet Data Analytics definieras enligt Runkler (2012) som tillämpning av analyser på stora datamängder i syfte att stödja viktiga beslutsfattanden (Runkler, 2012). Data Analytics är en generell term som används för alla typer av bearbetning av historisk data (Informatica, n.d.). I takt med att mängden data har vuxit har Data Analytics-termen utvecklats till att även inkludera Big Data (Informatica, n.d.). Därför kommer vi i denna uppsats använda litteratur som även behandlar Data Analytics i vissa fall.

I dagens konkurrensutsatta marknad har organisationer inom de flesta branscher ett behov av att använda sig av de enorma mängder data som kan användas för att erhålla ny kunskap (Dicuonzo, Galeone, Zappimbulso & Dell'Atti, 2019). Detta nya paradigm i den digitala eran kräver stöd av nya tekniker och kunskap inom programmering, modellering, rengöring och analysering av stora mängder data som extraheras från olika källor (Dicuonzo et al., 2019). Genom att tolka analysresultat är det möjligt att erhålla betydelsefull information som kan användas av beslutsfattare och i organisatoriska processer, vilket skapar värde för organisationen (Watson, 2014).

Likväl som Big Data Analytics medför stora möjligheter för organisationer, innebär det även en del utmaningar som organisationer måste ta i beaktning (Runkler, 2012). En av de största utmaningarna som upprepande förekommer i forskning är säkerställning av kvaliteten på datan som används vid analysen (Anandakumar, Arulmurugan & Suriya, 2019; Runkler, 2012). Utmaningar med datakvalitet är inte något nytt, men Big Data medför nya dimensioner till problemområdet i form av dataomvandling och hantering av enorma mängder data (Baesens, Bapna, Marsden, Vanthienen & Zhao, 2016). Om organisationer uppmärksammar och tar extra

hänsyn till utmaningar med Big Data Analytics kan detta innebära sparande av pengar och tid för organisationen (Gaikwad, Nale & Bachate, 2016).

1.2 Problemområde

Det finns flertal studier som belyser vilka utmaningar som kommer med att använda Big Data Analytics (t.ex. Anandakumar, Arulmurugan & Suriya, 2019; Marjani, Gani, Hashem, Siddiqa, Yaqoob, Nasaruddin & Karim, 2017; Gupta, Singhal & Garg, 2018). Trots detta saknas utförliga studier kring hur organisationer arbetar i praktiken för att bemöta dessa potentiella utmaningar. Big Data utgörs av enorma mängder data, vilket medför att all tillgänglig data inte är användbar för analys eller beslutsfattande (Gaikwad, Nale & Bachate, 2016). Enligt Gaikwad, Nale och Bachate (2016) är därför den största utmaningen med Big Data Analytics att utvinna värdefull data och för att lyckas med det menar Elgendy och Elragal (2014) att den data som analyseras måste vara av hög kvalitet. Det är således kritiskt för organisationer att förstå och kunna hantera problem associerade med datakvalitet (Batini & Scannapieca, 2006).

Trots att det finns framtagna Best Practises för datakvalitetsinitiativ är det inte ovanligt att förbättringar av datakvalitet misslyckas att uppnå ett långvarigt bidrag (Kokemüller, 2011). Enligt Kokemüller (2011) är de viktigaste faktorerna till ett framgångsrikt initiativ kända men han menar att de är baserade på anekdoter. I tillägg till detta belyser Tam & Kwan (2018) att datakvalitetens roll för Data Analytics inte uppmärksammats tillräckligt i dagens forskning. Med anledning av det har vi valt att studera hur organisationer arbetar i praktiken för att hantera utmaningar med datakvalitet som riskerar att uppstå i samband med Big Data Analytics.

1.3 Forskningsfråga och syfte

Vår forskningsfråga formuleras enligt följande: Hur hanterar organisationer utmaningar med datakvalitet inom Big Data Analytics?

Syftet med uppsatsen är att beskriva hur organisationer hanterar utmaningar med datakvalitet inom området Big Data Analytics.

1.4 Avgränsningar

Vi har valt att avgränsa oss till följande dimensioner inom datakvalitet: Information Completeness, Data Accuracy, Data Currency, Data Deduplication samt Data Consistency. Utöver utmaningarna inom ovan nämnda datakvalitetsdimensioner har vi identifierat andra utmaningar med datakvalitet inom Big Data Analytics. De utmaningar vi har kunnat urskilja från litteraturen och därmed fokuserar på i denna uppsats är: urval av Data Samples, felmarginaler, datarengöring och korrekt representation av data samt ledningens stöd och engagemang.

2 Litteraturgenomgång

I detta kapitel presenteras relevant teori som är nödvändig för att besvara vår frågeställning samt för att genomföra den empiriska studien. Kapitlet behandlar områdena Big Data, Big Data Analytics, datakvalitet, ETL-processen samt utmaningar med datakvalitet inom Big Data Analytics.

2.1 Big Data

Big Data utgörs av massiva mängder strukturerade och ostrukturerade datauppsättningar (Raheem, 2019). Strukturerad data syftar på data lagrad i databaser och ostrukturerad data avser data som inte finns lagrad i traditionella databaser bestående av rader och kolumner (Gaikwad, Nale & Bachate, 2016). Att endast definiera Big Data efter den stora mängden data den består av vore att begränsa begreppet (Raheem, 2019). Big Data handlar i första hand inte om själva datan i sig utan om det potentiella värde som kan utvinnas från den (Raheem, 2019).

Begreppet Big Data beskrivs i regel bestå av tre huvudfaktorer, de tre V:na, vilka är *Volume*, *Velocity* och *Variety* (Emmanuel & Stanier, 2016). Volume syftar på den enorma mängden av data som Big Data består av, Velocity avser den ständigt ökade hastigheten som data genereras och bearbetas på och Variety hänvisar till den breda variationen av olika datatyper och källor där data hämtas från (Emmanuel & Stanier, 2016). Den data som samlas in och lagras kan exempelvis komma från källor såsom sociala medier, maskiner, loggfiler, videos, texter, bilder och GPS:er (Watson, 2014). I tillägg till dessa tre V:n, som generellt används för att definiera de olika dimensionerna av Big Data, har IBM introducerat ytterligare två V:n, *Veracity* och *Value* (Yong, Shafei, Sian & Chua, 2019). Veracity är en datamängds kvalitetet och trovärdighet, det vill säga dess korrekthet samt datakällans pålitlighet (Yong et al. 2019). Value syftar på värdet som kan utvinnas av data, vilket potentiellt kan bidra till nya insikter och leda till bättre verksamhetsbeslut (Yong et al. 2019).

2.2 Big Data Analytics

Komplexiteten, variationen och den stora mängden data som Big Data medför har ställt nya krav på hantering och analys av data (Raheem, 2019). Detta har resulterat i att *Big Data Analytics* fått en central roll i dataanalysen och utvinningen av kunskap (Raheem, 2019). Big Data Analytics är en process som används för att analysera stora mängder data i syfte att upptäcka korrelationer och mönster som kan vara av värde och ligga till grund för framtida beslutsfattande (Shatnawi, Yassein, Abuein & Nsuir, 2019). *Data Analytics* är ett tvärvetenskapligt fält som använder aspekter från många andra vetenskapliga discipliner och metoder såsom statistik, mönsterigenkänning, Machine Learning (Runkler, 2012).

Big Data Analytics livscykeln kan delas in i sex olika faser (Shatnawi et al., 2019). Den första fasen är *Discovery* vilket innebär att ett problem identifieras och att det sedan görs upp en analytisk plan (Shatnawi et al., 2019). Den andra fasen är *Data Preparation* och innebär att kvaliteten på data bestäms för att kunna kontrollera om den är tillräckligt bra för att användas i en modell (Shatnawi et al., 2019). Vidare beskriver Shatnawi et al. (2019) att nästa fas heter *Modell Planning*, under denna fas undersöks det om förslaget av modell passar till analysen. Efter planeringen är nästa fas *Modell Building* och där kontrolleras det om den planerade modellen är tillräckligt kraftfull (Shatnawi et al., 2019). Fortsättningsvis förklarar författaren att nästa fas är *Operationalize*, under denna fas görs en slutlig dokumentation och sammanfattning samt leverans av projektkod och tekniska dokument (Shatnawi et al., 2019). Den sista fasen är *Communicate Results* som innefattar att ta reda på om resultaten har varit framgångsrika eller misslyckade (Shatnawi et al., 2019). Enligt Runkler (2012) kan Data Analytics projekt delas in i olika faser där data först väljs ut och utvärderas, sedan rengörs datan och filtreras, för att slutligen visualiseras och analyseras.

Inom Data Analytics går det att urskilja tre olika analysmetoder (Watson, 2014). *Descriptive Analytics* granskar data och information för att visa på vad som tidigare har hänt eller hur den nuvarande situationen ser ut (Watson, 2014). Vid Descriptive Analytics används exempelvis dashboards, scoreboards och datavisualisering (Watson, 2014). *Predictive Analytics* antyder på vad som kan komma att hända i framtiden med hjälp av prognostisering och statiska modeller såsom regressionsanalys och Machine Learning (Sivarajah, Kamal, Irani & Weerakkody, 2019). *Prescriptive Analytics* föreslår hur företaget bör gå tillväga för att optimera sina processer, förbättra sina servicenivåer och minska sina kostnader (Sivarajah et al., 2019). Sivarajah et al. (2019) menar på att det finns två ytterligare analysmetoder. Dessa är *Inquisitive Analytics* som hjälper företaget att förstå varför saker händer samt *Prescriptive Analytics* som refererar till förmågan att vidta åtgärder vid händelser som kan påverka företagets presentation negativt (Sivarajah et al., 2019).

2.3 Datakvalitet

Enligt Morbey (2013) definieras datakvalitet som den grad av uppfyllelse som har fastställts för data i ett specifikt syfte. I organisatoriska och individuella processer som är beroende av data har datakvaliteten blivit en avgörande faktor för kvaliteten av beslut och åtgärder (Zhang, Indulska & Sadiq, 2019). Om datakvaliteten är otillräcklig kommer vi exempelvis inte hitta korrekta svar på queries till databaser, oavsett hur skalbara och effektiva frågeställningsalgoritmerna är (Fan, 2015).

Tyvärr är data ofta inkonsekvent, felaktig, ofullständig, föråldrad och duplicerad (Fan, 2015). Faktum är att mer än 25% av all data i världens organisationer är felaktig och av dålig kvalitet (Fan, 2015). Dålig datakvalitet kan påverka analytiska resultat vilket i sin tur kan resultera i allvarliga förluster för organisationer (Zhang, Indulska & Sadiq, 2019). Datakvaliteten påverkar i sin tur även informationskvaliteten som endast kan bedömas utifrån kontexten där datan faktiskt används (Clarke & Taylor, 2018). Till följd av detta har flera initiativ lanserats, både i offentliga och privata sektorer, där datakvaliteten spelat en ledande roll (Zhang, Indulska & Sadiq, 2019). Exempelvis Data Quality Act, antagen av den amerikanska staten, samt Data Quality Assessment Methods and Tools (DatQAM) som stöttades av Europeiska kommissionen (Zhang, Indulska & Sadiq, 2019).

Inom datakvalitet finns det olika synsätt på vilka dimensioner och faktorer som utgör kvaliteten på data, där varje dimension av datakvalitet förser ett särskilt perspektiv (Heinrich, Hristova, Klier, Schiller & Szubartowicz, 2018). Det inkluderar giltighet, lämplig associering, lämplig signifikation, exakthet, precision och temporär tillämpbarhet (Clarke & Taylor, 2018). Det finns många olika identifierade dimensioner inom datakvalitet, men vissa omnämns oftare än andra (Sidi, Panahy, Affendey, Jabar, Ibrahim & Mustapha, 2012). De dimensioner som är vanligast att dela in datakvalitet i är *Information Completeness*, *Data Accuracy*, *Data Currency*, *Data Deduplication* och *Data Consistency* (Fan, 2015). Dessa dimensioner är särskilt viktiga vid enorma mängder av data och användning av *Data Warehouses* som ofta används som den primära datakällan vid Data Analytics och datarengöring (Blake & Mangiameli, 2011). Data Warehouses är centrala platser för datalagring som innehåller historisk data som en gång har genererats från ett eller flera transaktionssystem (Dupor & Jovanovic (2014).

2.3.1 Information Completeness

Information Completeness avser i vilken utsträckning data inte saknas samt att den har tillräcklig bredd och djup för den aktuella uppgiften (Pipino, Lee & Wang, 2002). Zhang, Indulska och Sadiq (2019) menar att Completeness syftar på att attribut som är nödvändiga för att ge en fullständig representation av verkligheten måste innehålla värden och får därför inte vara null. I linje med detta menar Fan (2015) att dimensionen avser om databasen har komplett och fullständig information för att svara på efterfrågade queries. Enligt Sidi et al. (2012) refererar Completeness till att informationen innehåller alla nödvändiga delar och värden som finns tillgängliga. En annan definition av Completeness enligt Bloland och MacNeil (2019) är att det är en indikator som reflekterar om all relevant data som behövs för att göra beslutsfattande, finns tillgänglig eller inte. Jesilevska (2017) namnger dimensionen som Data Completeness och menar på att det innebär data som uppfyller användarens behov.

2.3.2 Data Accuracy

Data Accuracy innebär till den omfattning som datan är korrekt, pålitlig och certifierad (Sidi et al., 2012). Data Accuracy syftar på att attributvärden måste vara exakta språkmässigt och ha korrekt detaljnivå (Zhang, Indulska & Sadiq, 2019). Fan (2015) menar på att Data Accuracy avser hur väl värden i en databas representerar de korrekta värdena av entiteterna som informationen i databasen syftar till. Data är exakt när datavärden överensstämmer med den riktiga världens värden (Yu, 2013). Vidare beskriver Jesilevska (2017) Data Accuracy som graden av reflektion utav den faktiska situationen.

2.3.3 Data Currency

Data Currency handlar om huruvida data är aktuell eller inte (Sidi et al., 2012). Vidare menar Sidi et al. (2012) att Currency beskriver tidpunkten då datan har blivit inlagd i en datakälla eller i ett Data Warehouse. Data kan anses vara aktuellt och relevant trots eventuella avvikelser orsakade av tidsrelaterade förändringar (Sidi et al., 2012). Enligt Yu (2013), Fan (2015) och Heinrich och Klier (2011) är Data Currency ofta refererad till Timeliness och dessa dimensioner kan därmed anses ha samma innebörd. Yu (2013) menar att tidsstämplar ofta är felaktiga eller inte tillgängliga i praktiken. Dessutom är datavärden ofta kopierade och importerade från olika källor vilket medför att de inte alltid har likadana format på sina tidsstämplar (Yu, 2013). Detta resulterar i att det blir mer utmanande att identifiera de “senaste” värdena på datan i databasen (Yu, 2013).

2.3.4 Data Deduplication

Data Deduplication avser till att identifiera en enda rad med korrekta värden och därmed ersätta dubbletter som refererar till samma riktiga entitet (Yu, 2013). Sidi et al. (2012) menar att Duplication är ett mått på oönskad duplicering inom i ett system för ett specifikt fält eller datauppsättning. Begreppet är också känt som record matching, record linkage, instansidentifiering samt objektidentifiering (Fan, 2015). Dimensionen har länge varit en utmaning och kan tänkas vara den datakvalitetsdimension som studeras och undersökts mest (Fan, 2015).

Behovet av Data Deduplication är tydligt inom exempelvis Data Quality Management och dataintegration (Fan, 2015). Dimensionen är särskilt betydande vid användning av Big Data (Fan, 2015). För att få praktisk användning av datan är det oftast nödvändigt att korrekt identifiera rader från olika källor som refererar till samma entitet, för att sedan förena datan och därmed förbättra informationen om entiteten (Fan, 2015). Fan (2015) menar på att det kan vara en svår uppgift i och med att datan kan vara smutsig samt att konflikter kan uppstå under föreningen och integreringen av datan, även fast datakällorna är relativt lika.

2.3.5 Data Consistency

Dimensionen Data Consistency innebär i vilken utsträckning data presenteras i samma format och är kompatibelt med tidigare data (Sidi et al., 2012). Vidare menar Sidi et al. (2012) att Consistency syftar på överträdelser av de semantiska regler som definierats för datasetet. Enligt Fan (2015) syftar dimensionen till att upptäcka fel i datan, inkonsekvenser och konflikter, oftast i typ av överträdelser mot databeroenden. Data Consistency omfattar två huvudsakliga uppgifter: att upptäcka och identifiera överträdelsen av databeroenden samt reparera och fixa felen som uppstått (Yu, 2013). Yu (2013) menar på att reparation och lagning av data ses som en av de vitala uppgifterna inom datarengöring.

2.4 ETL

ETL står för *Extract*, *Transform*, *Load* och är en komplicerad tidskrävande process som hanterar data med olika format och kvalitetssvårigheter (Souibgui, Atigui, Zammali, Cherfi & Yahia, 2019). ETL-processens syfte är att ta ut data från en källa och ladda in den till ett Data Warehouse samt att rengöra och transformera datan under vägen (Dupor & Jovanovic, 2014). Data Warehouses är centrala platser där enorma mängder data lagras från organisationers operativa databaser och externa källor (Dupor & Jovanovic, 2014; Souibgui et al., 2019). Souibgui et al. (2019) menar att noggrannheten och relevansen av Big Data Analytics är beroende av förmågan att inhämta data av hög kvalitet till Data Warehouses med hjälp av ETL-processen. Processen sker i tre steg: (1) extrahera data från datakällan, (2) transformera datan till korrekt format och (3) ladda in den slutgiltiga datan in i systemet (Souibgui et al., 2019).

Det finns många anledningar till behovet av en dataintegrationfas inom ett beslutsstödssystem (El Akkaoui, Zimányi, Mazón & Trujillo, 2011). Dessa skäl kan exempelvis vara heterogena format, att dataformat kan vara svårtolkade, att gamla system använder föråldrade databaser samt att datakällans struktur förändrats över tiden (Souibgui et al., 2019). Genom att integrera olika datakällor möjliggörs en fullständig och korrekt bild av organisationens operativa data (El Akkaoui et al., 2011). Enligt den studie som genomfördes 2019 av Souibgui et al. bidrar ETL-

processen till att förbättra datakvalitet, i viss utsträckning, i enlighet med de olika datakvalitetsdimensionerna. Exempel på några datakvalitetssvårigheter som ETL kan hjälpa till att åtgärda är brist på Integrity Constraints, saknade värden, variation av datatyper, namnkonflikter, repeterande data och inkonsekvent syntax (Souibgui et al., 2019).

2.5 Utmaningar med datakvalitet inom Big Data Analytics

Big Data Analytics kräver en hög kvalitet på data för att den ska vara användbar (Ijab, Surin & Nayan, 2019). Fernandes och Wagh (2019) anser att datakvalitet är stommen i alla analytiska lösningar. De menar därför att det är avgörande att säkerställa datakvalitet för Big Data innan denna data kan användas för analys, eftersom den annars riskerar att medföra felaktiga resultat (Fernandes & Wagh, 2019). Vissa forskare uppmanar av denna anledning till försiktighet vid användning av Big Data och Data Analytics (Clarke & Taylor, 2018). De menar att om datakvaliteten skulle vara bristfällig kommer Big Data Analytics endast kunna bidra med ett begränsat värde eller till och med orsaka negativa effekter på affärsresultat (Clarke & Taylor, 2018).

2.5.1 Urval av Data Samples

Fisher, DeLine, Czerwinski och Drucker (2012) menar att även fast organisationer har stor kvantitet av data innebär det inte att de sample de använder sig av är tillräckligt representativa för hela populationen. Det vill säga större är inte bättre (Fisher et al., 2012). Det betyder inte heller att man har en sanning av användarens verkliga behov eller motiv genom att ha stora mängder data om detta (Fisher et al., 2012). Oavsett hur stort ett dataset är så krävs det noggrann övervägning och tolkning för att få tillräcklig kvalitet på datan (Fisher et al., 2012). I motsats till detta beskriver viss litteratur att datakvalitet inte spelar någon roll vid användning av stora mängder data, som i fallet med Big Data (Clarke & Taylor, 2018). De menar att behovet av att identifiera ett noggrant urval av datan överträffas av möjligheten att ha tillgång till hela datasetet (Clarke & Taylor, 2018). Clarke och Taylor (2018) menar att om man har tillgång till hela populationen kan stora delar av datan vara rörig, men att i sådana volymer kan nya korrelationer hittas; med andra ord så blir kvantitet viktigare än kvalitet. Dock så finns det endast begränsade och väldigt specifika omständigheter då detta kan motiveras (Clarke & Taylor, 2018). I kontrast till detta menar Fernandes och Wagh (2019) att datakvalitet är grunden i alla analytiska lösningar vilket leder till att uppgiften med att säkerhetsställa kvaliteten på datan är mycket central.

2.5.2 Felmarginaler

Datakvalitet är ett multidimensionellt koncept (Beebe & Walz, 2005) och det finns de som menar att kvaliteten på data måste sättas i förhållande till vad den ska användas till (Baesens et al., 2016). Analyser inom olika områden kommer inte ha samma krav på exempelvis detaljnivå, syntax och kompatibilitet med tidigare data eller på hur aktuell datan är (Baesens et al., 2016). Med detta synsätt blir det således en stor utmaning att mäta datakvalitet, då samma data skulle kunna utvärderas olika beroende på vad den ska användas till (Baesens et al., 2016). Därför poängterar forskarna att data aldrig kommer att vara av perfekt kvalitet (Baesens et al., 2016). Det kommer med andra ord alltid finnas vissa felmarginaler och det är därför kritiskt att de felmarginaler som förekommer identifieras väl (Baesens et al., 2016). Det inkluderar även de effekter som kan uppstå till följd av felnivåerna och hur de påverkar resultaten vid en analys

(Baesens et al., 2016). Allt för ofta tycks organisationer betrakta investeringar i datakvalitet som för dyra eller svåra, vilket leder till misslyckande med att identifiera felmarginalerna och bedriva de potentiella vinsterna och fördelarna med Big Data Analytics (Baesens et al., 2016).

2.5.3 Datarengöring och korrekt representation av data

Datarengöring är ett samlingsnamn för tekniker som kan användas för att lösa problem med saknade värden, syntaktiska fel, ursprungligt låg kvalitet av data och avsaknad av metadata (Clarke & Taylor, 2018). Även om detta förbättrar datakvaliteten så uppstår det sällan ett tillstånd som kan beskrivas som ”ren data” (Baesens et al., 2016). De flesta rengöringsprocesser är manipulationer som baserats på statistiska analyser av datan (Clarke & Taylor, 2018). Många av de förändringar som görs inför dessutom andra fel (Clarke & Taylor, 2018). Vidare innebär rengöring av stora mängder data även att det lagras på flera olika platser och queries måste därför distribueras och skrivas så att det fungerar över ett helt nätverk (Fisher et al., 2012). Detta påverkar urvals- och rengöringsprocessen genom att potentiella fel och bias uppstår (Fisher et al., 2012).

Datarengöring för stora mängder data har betydande kostnader förknippade med sig (Beebe & Walz, 2005). Dålig kvalitet på data vid inmatning kräver ökad tid och ansträngning vid rengöringsprocessen utan några garantier för hög kvalitet (Lucas, Ishfaq & Raja, 2014). Kostnaden som är associerad med den manuella rengöringen tenderar ofta att vara linjär och ju större dataset desto högre kostnad för datarengöringen (Lucas, Ishfaq & Raja, 2014). Det finns både operativa kostnader som tillverkningsfel, långa ledtider samt medarbetares missnöje (Lucas, Ishfaq & Raja, 2014). Det finns dessutom strategiska kostnader som kan tillkomma såsom dålig planering, dåliga prispolicys samt minskad effektivitet (Lucas, Ishfaq & Raja, 2014). När man ska fatta beslut angående datakvalitet och Data Analytics är därför kostnad en viktig faktor att ta hänsyn till (Baesens et al., 2016). Utmaningen är att balansera kraven för datakvaliteten och resursbegränsningarna när det gäller tid, kostnad och expertis (Lucas, Ishfaq & Raja, 2014).

2.5.4 Ledningens stöd och engagemang

Det finns även utmaningar beträffande att få ledningens stöd och engagemang för datakvalitet och de analyser som genomförs för att dra beslutsgrundande slutsatser från detta (Baesens et al., 2016). Baesens et al. (2016) menar på att även organisationer med data av högsta kvalitet kommer misslyckas med att förbättra sitt datadrivna beslutsfattande om organisationens ledare inte litar på datan eller de analytiska tekniker som används. Att investera i datakvalitet är en långsiktig och krävande process med höga medföljande kostnader (Baesens et al., 2016). Därför krävs det att ledningen förstår vikten av hög kvalitet på data och litar på att datakvalitet är grunden för högt värderade och funktionella analysmodeller (Baesens et al., 2016). Även Beebe och Walz (2005) betonar vikten av ledningens stöd och engagemang i syfte att förbättra datakvaliteten.

Enligt Fass (2018) ligger Data Analytics i toppen av ledningens prioriteringar med syfte att förbättra organisationens kunskap och kapacitet. Det framkommer även att ett nyckelområde och en utmaning som finansledningar kämpar hårt för att hantera är just datakvaliteten inför analysprocessen (Fass, 2018). I motsats till detta menar Taskin, Pauleen, Scahill och Intezari (2019) att ungefär två tredjedelar av toppchefer inte har något förtroende för Big Data Analytics och förlitar sig hellre på deras egen intuition och erfarenhet vid beslutsfattande. Ledningen som inte förlitar sig på dataanalys bekymrar sig bland annat över låg datakvalitet, tillförlitlighet och

relevans av data samt datatillgång (Taskin et al., 2019). Detta innebär svårigheter att få stöd från toppledningen för att investera och använda Data Analytics (Taskin et al., 2019).

Analyser och rapporter som levereras ut till organisationen är helt beroende av kvaliteten och fullständigheten på den tillgängliga datan, vilket gör att ledningens investeringar i Data Analytics och datakvalitet är ännu mer kritisk (Fass, 2018). Det är därför nödvändigt och viktigt att utbilda toppledning och chefer inom Data Analytics och att använda data för förbättrat beslutsfattande (Taskin et al., 2019). För att övervinna ledningens motstånd krävs det en öppen kommunikation om detta inom organisationen, en datadriven kultur och utbildning inom datadrivna analyser och teknisk förståelse för analys hos ledningen (Taskin et al., 2019).

2.6 Sammanställning av litteraturgenomgång

Tabell 2.1: Översikt av teori

Område	Aspekter	Litteratur
Big Data	● Allmänt om Big Data	(Raheem, 2019); (Gaikwad, Nale & Bachate, 2016); (Emmanuel & Stanier, 2016); (Watson, 2014); (Yong et al. 2019);
Big Data Analytics	● Allmänt om Big Data Analytics	(Raheem, 2019); (Shatnawi et al., 2019); (Runkler, 2012); (Watson, 2014); (Sivarajah et al., 2019);
Datakvalitet	● Allmänt om datakvalitet	(Morbey, 2013); (Zhang, Indulska & Sadiq, 2019); (Fan, 2015); (Clarke & Taylor, 2018); (Heinrich et al., 2018); (Sidi et al., 2012); (Blake & Mangiameli, 2011); (Dupor & Jovanovic (2014);
	● Datakvalitets dimensioner - Information Completeness - Data Accuracy - Data Currency - Data Deduplication - Data Consistency	(Pipino et al., 2002); (Zhang, Indulska & Sadiq, 2019); (Fan, 2015); (Sidi et al., 2012); (Bloland & MacNeil, 2019); (Jesilevska, 2017); (Yu, 2013); (Heinrich & Klier, 2011);
	● ETL	(Souibgui et al., 2019); (Dupor & Jovanovic, 2014); (El Akkaoui et al., 2011);

Utmaningar inom datakvalitet i Big Data Analytics	• Allmänt om utmaningar inom datakvalitet i Big Data Analytics • Urval av data samples • Felmarginaler • Datarengöring och korrekt representation av data • Ledningens stöd och engagemang	(Ijab, Surin & Nayan, 2019); (Fernandes & Wagh, 2019); (Clarke & Taylor, 2018); (Clarke & Taylor, 2018); (Fernandes & Wagh, 2019); (Fisher et al., 2012); (Beebe & Walz, 2005); (Baesens et al., 2016); (Clarke & Taylor, 2018); (Baesens et al., 2016); (Fisher et al., 2012); (Beebe & Walz, 2005); (Lucas, Ishfaq & Raja, 2014); (Baesens et al., 2016); (Beebe & Walz, 2005); (Fass, 2018); (Taskin et al., 2019);

3 Metod

I detta kapitel kommer vi inledningsvis redogöra för vilken metod som använts för insamling av data och vårt urval av respondenter. Vi kommer därefter ingående beskriva hur intervjuerna genomförts och avslutningsvis diskutera aspekter som validitet och etik.

3.1 Metodval

Vår uppsats har en beskrivande frågeställning med syftet att beskriva hur utmaningar med datakvalitet hanteras inom Big Data Analytics. Vi har därför valt att genomföra en kvalitativ studie bestående av fem semistrukturerade intervjuer för att besvara vår forskningsfråga. Jacobsen (2002) menar att kvalitativa metoder är fördelaktiga att använda om avsikten är att undersöka befintliga teorier eller skapa en bättre förståelse inom ett område. Vi anser att intervjuer är ett lämpligt format för kvalitativ datainsamling då det enligt Jacobsen (2002) tillåter den som intervjuas att förmedla sina perspektiv och dela med sig av sina erfarenheter mer fritt. Det erbjuder större flexibilitet genom möjlighet till att ställa spontana följdfrågor som kan hjälpa till att uppnå mer uttömmande och detaljerade svar (Jacobsen, 2002).

Kvalitativa metoder har dock mottagit kritik för att de inte behandlar tillräckligt mycket data i sina analyser (Oates, 2006). Vi har därför försökt reflektera över kopplingar mellan vald teori och vår empiri för att skapa en mer djupgående analys, vilket Ryen (2004) menar är en av de främsta fördelarna med användning av kvalitativa studier. Enligt Alvehus (2013) används kvalitativa metoder för tolkande forskning, där data samlas in och skrivs i löpande text för att sedan tolkas. Kvalitativa studier fokuserar på data i form av ord vid insamling och analys (Bryman, 2012) och enligt Jacobsen (2002) är vi i regel mer mottagliga för ny information vid användning av kvalitativa studier. Med tanke på att vår insamlade data ska användas för att beskriva hur organisationer hanterar utmaningar med datakvalitet inom Big Data Analytics ser ut valde vi därför ett kvalitativt tillvägagångssätt.

3.2 Urval

Enligt Jacobsen (2002) bör ett urval styras utifrån syftet med studien. Kompetensen hos de intervjupersoner som väljs ut behöver därför baseras på den information som forskningsfrågan ska besvara. För att kunna besvara vår forskningsfråga ansåg vi att de personer som skulle intervjuas behövde ha erfarenhet av att arbeta med Big Data Analytics. Det blev därav naturligt att vi riktade in oss mot personer som arbetar i roller såsom Data Scientist eller Data Analyst.

Under det första urvalet som gjordes använde vi vårt kontaktnät för att komma i kontakt med personer som kunde tänka sig delta i en intervju och som passade in på vårt urvalskriterium. Detta kallas enligt Jacobsen (2002) för ett bekvämlighetsurval. Under den andra urvalsprocessen mailade vi till olika större organisationer som arbetar med Big Data Analytics antingen internt

med sin egen data eller utifrån ett konsultperspektiv där de analyserar sina kunders data. På så vis lyckades vi komma i kontakt med personer inom dessa organisationer som arbetade i en lämplig roll för vår studie. Nedan har vi sammanställt en tabell över de företag och respondenter som vi intervjuat.

Tabell 3.1: Respondenter

Företag	Respondent	Roll och erfarenhet	Datum	Intervjutid	Appendix
F1	R1	Head of Data Science and AI	2020-04-29	55 minuter	A
F2	R2	Data Scientist konsult	2020-04-29	40 minuter	B
F3	R3 & R4	Management Consultants *Erfarenhet inom AI, Machine Learning och Analytics*	2020-05-05	44 minuter	C
F3	R5 & R6	Data Analysts *Under avdelningen för revision*	2020-05-05	71 minuter	D
F4	R7	Senior Advisory *Under avdelningen Governance Risk and Compliance*	2020-05-07	42 minuter	E

3.3 Intervju

Med tanke på uppsatsens beskrivande frågeställning har vi valt att ställa öppna och strukturerade intervjufrågor med möjlighet att ställa följdfrågor. Att använda denna typ av semistrukturerad intervju öppnar även upp för diskussioner och möjlighet att ändra om ordningen på frågor (Jacobsen, 2002). Detta leder till en mer detaljerad beskrivning där respondenten ges möjlighet att uttrycka sitt perspektiv och utförligt förklara hur det ser ut specifikt inom deras organisation (Jacobsen, 2002). Detta är användbart inom vår studie då alla organisationerna troligtvis inte drabbas av samma utmaningar med datakvalitet, samt att de hanterar dessa på olika sätt. Vi vill därför ha en så utförlig och detaljerad beskrivning som möjligt av hur var och en av respondenterna hanterar sina utmaningar.

Viktigt att tänka på vid semistrukturerade intervjuer är att lyssna noggrant, inte avbryta intervjuobjektet men samtidigt jobba med följdfrågor för att kunna samla in data som bidrar till att besvara frågeställningen (Alvehus, 2013). Det ger oss större ansvar som intervjuare att se till så att innehållet är relevant och att intervjun hålls inom ämnet (Alvehus, 2013). Detta hanterade vi genom att ha förberedda potentiella följdfrågor som vi tog fram i syfte att antingen leda tillbaka intervjun inom rätt ämne eller att få fram mer detaljerad information om respondentens svar. Vid utformningen av frågor såg vi även till att ställa öppna frågor som inte ger ledande svar.

Jacobsen (2002) menar att insamling av kvalitativ data är väldigt tidskrävande och därför rekommenderar han att endast genomföra ett begränsat antal intervjuer. Av denna anledning var det viktigt för oss att genomföra ett par längre intervjuer med en till två deltagare för att lyckas samla in vad vi ansåg vara tillräckligt mycket data för vår diskussion. Vidare var det tidskrävande att först hitta respondenter som jobbar med Big Data Analytics och som dessutom var insatta i

företagets processer med datakvalitet. Vi påbörjade denna process redan i uppstarten när frågeställningen var färdigställd, vilket visade sig vara tacksamt då det var en tidskrävande process.

Det framgår inte i litteraturen hur organisationer bör gå tillväga för att hantera utmaningar med datakvalitet ur ett Big Data Analytics-perspektiv och vi har därför varit begränsade i den jämförelse vi kunnat göra mellan teori och praktik i detta avseende. Av denna anledning identifierade vi utmaningar i litteraturen och frågade sedan respondenterna om dessa var förekommande i deras arbete med Big Data Analytics och hur de i sådana fall hanterades.

3.3.1 Genomförande av intervju

De fem intervjuerna genomfördes genom videosamtal på Microsoft Teams. Anledningarna till att intervjuerna genomfördes på distans var dels på grund av rådande situation med Covid-19 och dels för att majoriteten av respondenterna befinner sig utanför Skåne. Enligt Jacobsen (2002) påverkas personer även av den omgivande miljön vid en intervju, vilket kan påverka beteende och hur personen svarar på intervjufrågorna. Tack vare att vi använde oss av videosamtal befann sig respondenterna i sin naturliga miljö vilket ger mer validitet på svaren. En nackdel med detta kan emellertid vara att vi missar hur respondenten beter sig (Jacobsen, 2002). Exempelvis blir det svårt att observera kroppsrörelser och ansiktsuttryck. Men lyckligt nog använde fyra av fem intervjuer videosamtal så att vi även kunde tolka deras ansiktsuttryck och kroppsrörelser.

Under två av intervjuerna var det två respondenter som representerade företaget. De andra tre intervjuerna var individuella intervjuer. Det blev en bra blandning av olika perspektiv från intervjuerna med två respondenter, men även lite mer djupdykning under de individuella intervjuerna. Inför intervjuerna hade vi skapat en intervjuguide som var baserad på vår litteraturgenomgång för att ha som utgångspunkt. Intervjuguiden är indelad i sju olika delar, där varje del berörs ett visst område eller utmaning. De sju olika delarna innefattar: (1) Introduktion och Bakgrund, (2) De olika dimensionerna av datakvalitet, (3) Användning av ETL, (4) Urval av Data Samples, (5) Felmarginaler, (6) Datarengöring och korrekt representation av data och (7) Ledningens stöd och engagemang.

Intervjuerna inleddes med en presentation om oss, vad uppsatsen handlar om samt en förklaring av hur intervjun skulle gå till. Vi frågade även om godkännande av inspelning av intervjun samt tillåtelse om att publicera den i vår kandidatuppsats. I tillägg till detta beskrev vi även de etiska aspekterna. Vi gick igenom intervjuguiden uppifrån och ner med eventuella följdfrågor under intervjuns gång. I slutet ställde vi en öppen fråga om det var något annat respondenten ville tillägga eller tyckte att vi hade missat. Intervjuerna tog mellan 40-70 min att genomföra.

Efter att alla fem intervjuer var utförda så transkriberade vi dessa för att samla in all information som respondenterna gav oss. Enligt Bryman (2013) är denna metod bra då alla deltagare i intervjun kan fokusera fullt på samtalet och inte bekymra sig över att missa viktiga bidrag eller stressa med att anteckna för hand. Genom att transkribera intervjuerna blev det även lättare för oss att senare sammanställa och analysera vad varje respondent berättat samt att ställa detta mot varandra. Analysen av sammanställningen gjordes genom att jämföra svaren och de perspektiv vi fick från intervjuerna med den litteratur vi har valt.

Tabell 3.2: Intervjuguide

Område	Aspekt	Exempelfrågor	Intervjufrågor
Bakgrund	● Allmänt om respondenten	- I vilken roll arbetar du som?	1, 1a, 1b
Datakvalitet	● Datakvalitetens dimensioner - Information Completeness - Data Accuracy - Data Currency - Data Deduplication - Data Consistency	- Vilka utmaningar finns det inom Information Completeness? Och hur hanterar ni dessa?	2, 2a 3, 3a 4, 4a 5, 5a 6, 6a
	● ETL	- Använder ni er utav ETL för att säkerställa datakvaliteten inom Big Data Analytics?	7, 7a, 7b
Utmaningar inom datakvalitet inom Big Data Analytics	● Urval av data samples	- Hur hanterar ni utmaningen med att välja ut data samples som ger en rättvis representation av datan?	8, 8a
	● Felmarginaler	- Hur hanterar ni dessa felmarginaler?	9, 10, 11
	● Datarengöring och korrekt representation av data	- Vilka utmaningar anser du att det finns med datarengöring? Hur hanterar ni dessa utmaningar?	12, 13, 13a, 14, 15
	● Ledningens stöd och engagemang	- Hur arbetar ledningen för att ge stöd till ert arbete?	16, 16a, 16b
Övrigt		- Finns det några fler utmaningar med datakvalitet inom Big Data Analytics som vi inte redan nämnt? Och hur hanterar ni dessa?	17, 17a

3.4 Etik

Inför intervjuerna fick respondenterna inte ta del av intervjufrågorna då vi ville ha mer spontana svar där respondenterna utgår från deras egna perspektiv och erfarenheter. Dock skrevs det ihop ett mail som vi skickade ut där vi beskrev vilka områden intervjun kommer att behandla så att respondenterna har möjlighet att förbereda sig. Som tidigare nämnts så bad vi om lov att spela in intervjuerna så att vi i senare skede kunde transkribera, vilket alla respondenter tillät.

Alla sju respondenter samt respektive organisation gavs möjligheten att vara anonyma. Oates (2005) menar att det är viktigt för att skydda respondenterna mot repressalier om de exempelvis skulle dela med sig av känslig information. Även Jacobsen (2002) beskriver vikten av detta då risken är större att respondenterna nekar att deltaga om de ska bli refererade till med namn. Även fast några av respondenterna godkände att inte vara anonyma så valde vi att både respondenter och dess företag ska hållas anonymt så att alla behandlas på samma sätt. Det som beskrivs är vilken arbetstitel och typ av företag respondenterna har och jobbar på.

Respondenterna delges information om studiens syfte, vilka vi är samt hur vi kommer att använda intervjuerna i uppsatsen. Alla sju försäkras om att det material som samlas in endast kommer att användas för att besvara studiens syfte. Genom att göra detta minskar osäkerheten kring respondenternas inställning till intervjun och undran om hur deras åsikter kommer att bearbetas (Jacobsen, 2002). Innan intervjun påbörjades så erbjöds respondenterna även att få transkriberingen skickad till sig så att de hade möjlighet att läsa igenom och möjligtvis radera viss information som sades i intervjun. Vi upplevde att respondenterna genast kände sig mer bekväma efter att detta sagts och det är även något som Oates (2005) understryker.

3.5 Validitet och reliabilitet

Validitet och reliabilitet handlar om att studien mäter det som är relevant och att den är tillförlitlig, vilket är viktiga kriterier vid utförandet av en kvalitativ studie (Jacobsen, 2002). Respondenterna vi har valt arbetar i huvudsak med Big Data Analytics och tampas med utmaningar inom datakvalitet dagligen. Respondenternas svar är därför både relevanta och giltiga att basera vår studie på. Den semistrukturerade kvalitativa intervjun bidrog även till validitet då vi kunde vidareutveckla svar och erhålla mer detaljerade svar utifrån varje respondents perspektiv samt förstå hur just deras organisation hanterar utmaningarna. Jacobsen (2002) styrker även detta då han menar att individuella intervjuer bidrar till validitet då personliga erfarenheter och åsikter från respondenten kan bidra till forskningsområdet som studien undersöker. Intervjufrågorna är direkt baserade på vår forskningsfråga och litteraturgenomgång vilket gör att vi verkligen får svar på det vi vill samt uppfyller uppsatsens syfte. Litteraturen är i sin tur även noga granskad och utvald där källor har jämförts och ställts emot varandra.

Vi uppfyller även reliabilitet genom bland annat att spela in intervjuerna samt att transkribera ordagrant, vilket leder till att all information från intervjuerna fångas upp och tas med i analysen. Inför varje intervju förklarades syftet med studien för varje respondent på samma sätt, så att varje person fick samma information och grund att utgå ifrån. Att vi efterhand även hade möjlighet att kontakta respondenterna för eventuella förtydliganden stärkte även reliabiliteten då inga antaganden behöver göras. Det som dock kan tas i beaktning vid användning av slutsatsen är antalet intervjuer som empirin är begränsad av. Trots att det är fem djupgående intervjuer med sju personer totalt bör resultaten ses utifrån varje situation och inte nödvändigtvis generaliseras.

4 Resultat

Under detta kapitel kommer det empiriska resultatet att presenteras. Datan samlades in under våra fem intervjuer. Resultatet baseras på dessa intervjuer och är strukturerat utifrån de utmaningar vi tidigare har presenterat i litteraturgenomgången. Kapitlet kommer inledas med en definitionslista som beskriver termer respondenterna använt under intervjuerna som inte förklarats tidigare i teorin.

Tabell 4.1: Definitioner för resultatet

Begrepp	**Definition**
Constraint	En Constraint är en regel som kan används för att optimera data (IBM, n.d.). Det finns olika varianter av Constraints som reglerar olika delar av datans format (IBM, n.d.).
Data Warehouse	Data Warehouses är centrala platser som lagrar enorma mängder historisk data som en gång har genererats från ett eller flera system (Dupor & Jovanovic, 2014).
Ensemble Model	Ensemble metoder är tekniker som skapar flera analysmodeller och sedan sammanslår dessa modeller till ett mer träffsäkert resultat (Demir, n.d.).
Histogram	Ett histogram är en form av diagram som grupperar numeriska data i olika sektioner och som presenterar dessa sektioner i olika kolumner (Google, n.d.). Det kan användas för att visa spridningen i ett dataset (Google, n.d.).
Outlier	Inom statistik är en Outlier en datapunkt som skiljer sig väsentligt ifrån resterande datapunkter (Dougherty, 2016).
RowCount	RowCount är ett kommando som används för att räkna antalet datarader (SmartBear, 2020).
Single Point of Truth/ Single Source of Truth	En gemensam databasstrategi med syfte att uppnå en enda källa för all design- och informationshantering av databeroenden, relationer och förändringar (Chown, 2018).
Query	Ett sätt att hämta en delmängd information från en databas (Churcher, 2008).

4.1 Information Completeness

Samtliga respondenter var eniga om att de upplevt utmaningar inom dimensionen Information Completeness i deras arbete. De tog upp olika branschspecifika exempel på utmaningar de stött på, men gemensamt för dessa var att alla relaterade specifikt till problemet med avsaknad av data. Med det sagt förekom det även andra exempel på utmaningar med samband till datas detaljnivå. Det var dock inte alla respondenter som ansåg att detta som en stor utmaning för dem. Respondent 1 (R1) som arbetar i rollen som IT-chef beskrev utmaningen med Information Completeness enligt följande:

> … vi känner absolut igen problemet vi stöter på det ibland, men jag skulle säga att det inte är ett av våra största datakvalitetsproblem för vi är lyckligt lottade att ha väldigt mycket data på F1 och vi har ju också väldigt detaljerad data…
> (Bilaga 2, rad 10)

Vidare beskriver R1 att anledningen till att de lyckas bra med hanteringen på denna punkt var dels för att transaktionsdata inom dagligvaruhandel är väldigt detaljerad i sig och dels för att en väldigt stor andel av deras kunder väljer att identifiera sig med sitt kundbonus-kort när de handlar. Hen förklarar sedan att om de saknar data så beror det främst på att de inte börjat lagra den på ett lämpligt eller strukturerat sätt ännu. För att hantera utmaningen med ostrukturerad data, så beskriver R1 att Data Scientists får i ansvar att extrahera och sammanställa den data som behövs i ett strukturerat format. Gällande den data som aldrig lagrats eller raderats finns det ingen möjlighet att åtgärda problemet i efterhand enligt R1, utan då bestämmer de istället huruvida de vill börja lagra denna typ av data för framtiden och avvaktar sedan i väntan på att de ska få tillräckligt mycket data till en analys.

Respondent 2 (R2) som arbetar som konsult åt ett förpackningsindustribolag beskrev att hen arbetat i ett analysprojekt där de ansåg att den data som skulle användas hade bristande bredd och djup för uppgiften. De hanterade då denna utmaning genom att använda sig av Ensemble modeller som gav dem olika prediktioner som de sedan använde för att ta fram ett generellt medelvärde för utfallen.

Respondent 3 och 4 (R3 och R4) som har erfarenhet av att arbeta med AI, Machine Learning och Data Analytics på en stor revisionsbyrå berättar att det vanligaste området som de tillämpar Data Science på är inom ekonomi, såsom att tolka kvitton och fakturor. R3 betonar att Finansinspektionen ställer strikta krav inom detta område, vilket innebär att de krävs fullständig data och att värden inte får saknas. Således menar R3 och R4 att de i normala fall inte har problem med denna utmaning då de i regel får fullständig strukturerad data.

Senare berättar de dock att de vanligtvis hanterar utmaningar med Information Completeness genom att sätta ”upp en threshold på sina modeller eller sitt hjälpmedel i form av AI” (Bilaga 4, Rad 9) som hjälper till att kontrollera osäkerheter i deras data. De förklarar att själva åtgärden som tas vid saknad data kan variera beroende på kostnaderna förknippade med felet ”man får göra en bedömning av hur dyrt felet faktiskt skulle kunna vara” (Bilaga 4, Rad 9). Om det är oväsentlig data kan de låta det vara, men om det är data av större betydelse som skulle kunna orsaka framtida kostnader kontrolleras uppgifterna och hanteras manuellt. Detta synsätt delas även av Respondent 5, som uttryckte att beroende på vad som efterfrågades och på vilka krav

som ställdes ”trying to ascertain if the data is complete or not, sometimes there will be an allowed sort of tolerance or difference” (Bilaga 5, Rad 18).

Både Respondent 5, 6 och 7 (R5, R6 & R7) som alla arbetar med Big Data Analytics på revisionsbyråer, upplever att problem vid kundernas extrahering av data är en vanlig orsak till att den blir ofullständig och att det kan leda till saknade värden. De ser det därför som en utmaning inom dimensionen Information Completeness. För att hantera detta beskriver R6 att de gör en extrahering och jämför med exempelvis en aggregerad rapport från samma system för att säkerställa att datasetet är komplett eller att de kontrollerar rapportens logik för att säkerställa att de är uppbyggda på samma vis.

Detta arbetssätt var även R5 familjär med som beskrev processen på ett likartat sätt. I tillägg beskriver R7 att de i vissa fall hanterar utmaningen med dataextraheringen genom att de går igenom extraheringsprocessen tillsammans med kunden alternativt gör den åt dem. Vidare berättade R5 att det kan vara svårare att kontrollera fullständigheten i icke-finansiella data, som exempelvis data i en statisk lista med leverantörsuppgifter. Det framgår dock att de även här kan använda sig av en RowCount för att kontrollera att alla rader i listan finns med.

4.2 Data Accuracy

Alla respondenter ansåg att utmaningen med Data Accuracy är något dem stöter på och hanterar ofta i sitt arbete. Det skiljer sig mycket på hur detta sker beroende på både organisation och person. Många står ofta i valet om de ska korrigera de felen som finns i datan eller låta det vara, exempelvis lämna en rad blank. R5 berättar om ett projekt då de tog datan och la upp en interaktiv dashboard så att klienten kunde se hur datakvaliteten såg ut. De flesta fel bestod av problem inom Data Accuracy då det ofta är klientens kunder som manuellt fyller i datan. Exempel på fel var förnamn som endast var en bokstav och tomma fält som var frivilliga att fylla i, exempelvis telefonnummer. R5 förklarar hur de hanterar detta:

> Because we had access to the raw data we could pull out specific examples, we could adapt. The client could then go back and recontact customers if they wanted to. We also, I suppose depending on what they were specifically looking for, we could refine the analytics so we didn’t treat... we could treat a blank entry as not being inaccurate data as such it was just missing (Bilaga 5, Rad 22).

Hen förklarar vidare att den klienten valde att rätta till felen då det ställs hårda krav på att data de lagrar om personer ska vara korrekt och up-to-date. Även R2 förklarar att de också försöker att rätta till de fel som finns i datan. I första hand pratar hen med en expert som kan datan så att man tillsammans kan fylla i luckor och ändra felaktiga värden. R2 förklarar också att de kan använda sig av kolumner som heter Previous Value, så man lätt kan återskapa den förra raden med hjälp av den senare raden och leta efter ledtrådar för att återskapa det som saknas eller är fel.

Till skillnad från detta berättar R1 att det svåraste är att identifiera själva felen i datan. Hen beskriver att det är nästan omöjligt att ta reda på om det inte är uppenbara fel som när man förväntar sig en siffra och det står text istället. Men i fall där det står en sjua istället för en nia är jättesvårt att märka. Hen berättar att de främst använder två metoder för att hantera detta, där den första beskrivs på följande vis:

> Men ett alternativ är helt enkelt att exkludera den delen av data i analysen. Och då också såklart tydligt beskriva det när man presenterar sin analys liksom att den här datan under den här perioden är inte liksom, ingår inte i modellen exempel på grund av det här och det här (Bilaga 2, Rad 18).

Alternativt hanteras det enligt beskrivningen nedan:

> … ett annat alternativ är att man inkluderar det, men att man då är extremt noga med att dokumentera. Och också när man presenterar data att man då pressar in en blasklapp i att den som tolkar den här analysen är medveten om att det finns felaktigheter (Bilaga 2, Rad 18).

R3 är också inne på samma spår som R1, alltså att friskriva sig och förklara att datan rör sig i ett väldigt brett spann och att inte garantera att allting stämmer när de hämtar stora mängder data från öppna datakällor. Däremot om de ska använda en kunds egna data så hade hen försökt säkerställa att datan stämmer samt att göra en prognos och förklaringar redan innan utvecklingen påbörjas.

4.3 Data Currency

I princip alla av respondenterna menar på att Data Currency, aktuell data, inte är en större utmaning för dem. R3 och R4 påpekar att de inte har någon erfarenhet utav problemet eftersom de oftast arbetar med realtidsdata. R4 underströk dock betydelsen av att kontrollera att datan är så pass aktuell att den är relevant att använda. Både R4 och R6 beskriver att detta stäms av genom kontroller under första genomgången av rådatan. Även R5, R6, och R7 menar på att Data Currency inte är ett problem för dem, eftersom de inom revision arbetar med färsk data ett år i taget. R1 berättar att de använder sig utav tidsstämplar. Hen menar att hen inte ser någon utmaning med att få tidsstämplarna överensstämmande i och med att de oftast inte används med en precision högre än en timma.

R6 beskriver ett exempel på en Data Currency utmaning då de hade flera leverantörer med liknande namn vilket indikerade på att datan inte var så aktuell. Hen hanterade detta genom att kolla på giltighetstiden på datan, som då kan antyda på vilken data som är mest up to date. R5 berättar om ytterligare ett exempel där deras klienter har haft interna CRM-system för finansiell Treasure Operations som har hämtat valutakurser, men samtidigt haft externa finansiella instrument vars värde var baserade på bankers valutakurser. Respondenten menar att det då uppstår skillnader i klientens redovisning just för att datan inte är “up to date”. Här gäller det då för klienterna att värdera hur stora skillnaderna är i redovisningen och om det är rimligt. Klienterna får då hantera denna svårighet genom att själva avgöra om de manuellt behöver rätta till felet.

R1 tror att det finns framtida utmaningar med aktuell data på F1, då de skulle vilja börja ägna sig mer åt prediktioner och analyser i realtid. Detta är en utmaning eftersom de idag använder sig av dygnsladdningar av data. R1 menar på att ett infrastruktursarbete skulle behöva genomföras där man kopplar in realdata till deras plattform för att bemöta denna utmaning.

En utmaning inom Data Currency som R2 stött på var när hen skulle börja träna en modell i syfte i att se om ett projekt i framtiden kan komma att överskrida en viss budget. Hen åsyftar då på att man kollar på tidigare avslutade projekt för att kunna få in samples till modellen. I ett projekt kollade R2 på projekt som var upp till tio år gamla, för att kunna få in så många samples som

möjligt att arbeta med. Problemet var då att projekten inte var relevanta längre, eftersom fel som förekom för tio år sedan är fel som inte skulle förekomma idag. R2 och hens team hanterade då denna utmaning genom att inte använda sig utav de äldre projekten, utan nöjde sig med de samples de hade.

4.4 Data Deduplication

Utmaningen med dubbletter i datan är något alla sju respondenter känner till och är medvetna om, men är enligt respondenterna inte en särskilt vanlig förekommande utmaning. Varken R1, R2, R5 eller R7 tycker att det är något stort problem. Om det någon gång skulle ske så hanterar R1 utmaningen genom att använda Constraint-kontroller i databaserna och Data Warehouses för att fånga upp dessa dubbletter. R3 och R4 tyckte att utmaningen med Data Deduplication var mer vanligt förekommande i deras arbete och R3 hanterar detta genom följande:

> Alltså antingen sorterar du ut det i ett ganska tidigt skede genom queries där du kan ta SELECT DISTINCT till exempel [...] Men sen också att jag när jag bearbetar datan så brukar jag föra register på liksom vad datan innehållit. Att jag kan ta all data som jag extraherar till exempel från ett dokument och sen ta den och lägga den någonstans och sen så jämför jag den hela tiden så att datan är unik (Bilaga 4, Rad 21).

R5 som inte brukar mötas av denna utmaning anser att det i sådant fall är ett extraheringsfel när datan hämtas ut, men beroende på vilken nivå av duplication så brukar det gå att själv fixa genom att koda och gruppera raderna till en. R6 som arbetar på samma företag anser däremot att detta problem är desto vanligare i hens arbete. Det händer när ’’the lowest common denominator is not unique on every row so when you kind of aggregate the dataset you get a lot of duplicates” (Bilaga 5, Rad 37). R6 menar att när du joinar tabeller så dupliceras raderna då det är två rader av samma i en av tabellerna men inte i den andra. Hen hanterar då detta på samma sätt som R5 beskrev, plus att enkelt kolla om det finns dubbletter i datasetet genom att räkna antalet rader som har samma värden och se efter om det finns dubbletter eller inte.

Alla respondenter verkar vara överens om att utmaningen med Data Deduplication främst uppstår när man för samman data från olika system och källor. Exempelvis menar R1 att när vi kopplar ihop data från två olika bolag så kan man få att en och samma kund har ett id i ett av bolagen och ett annat id i det andra bolaget. Majoriteten av respondenterna nämner även att det nästan alltid krävs en manuell handpåläggning vid dessa situationer och de benämner vikten att ha en unik identifierare av varje rad. Exempelvis beskriver R4 att:

> … i de fallen man har hämtat data ifrån olika källor som kan ha konflikterad information så måste man ju på något sätt använda någon slags unik identifierare för varje rad när man för samman datan till en gemensam databas… (Bilaga 4, Rad 23)

Hen beskriver vidare att de använder verktyg som Google Clouds plattform för att hantera det. R5 menar också att utmaningen med dubbletter uppstår när de inte har någon unik identifierare av datan. Alltid när man begär data så utgår man från ett unikt id, och om det inte finns så går det att skapa på eget vis. Ett exempel är då som också tidigare var ett förslag från R6 att stapla de olika fälten ihop och så att varje rad är unik och har ett unikt id. Vidare beskriver R7 att detta vanligtvis hanteras genom kraftfulla verktyg som hanterar detta inom hens organisation, eller eventuellt genom att ta bort en av kolumnerna där dubbletterna finns. I sådant fall poängterar R2

att det är viktigt att man pratar med någon som kan datan väldigt bra, så att man förstår vilken av dubbletterna som ska behållas.

4.5 Data Consistency

Majoriteten av respondenterna (R3, R4, R5, R6 samt R7) ansåg att Data Consistency är en svårighet hos dem. Samtliga av dessa respondenter arbetar helt eller till viss del i rollen som konsult. Dessa respondenter hämtar in andra organisationers data där alla har olika system, vilket leder till att exporten av information kan se olika ut. R7 menar att det är svårt att påverka hur exporten av information från andra organisationer ser ut. Det enda hen kan göra är att ge rekommendationer om hur hen önskar att datasetet ska se ut för att analyserna ska ske så effektivt som möjligt. Vidare berättar R7 att om hen får ett dataset i fel format så hanteras problemet genom att tvätta och transformera datan till korrekt format. Efter detta utförs kontroller på den omvandlade datan för att verifiera att det är samma data som innan rengöringen och transformationen genomfördes.

Vid bokföring använder sig R7 av standardiserade mallar. Fortsättningsvis menar R5 och R6 att de även använder sig av standardiserade mallar. Om datan inte är kompatibel till dessa bearbetas den däremot om tills datan kan användas till analys. I ett exempel hade R5 erhållit två set av samma transaktionsdata från två olika personer, där ena datasetet innehöll fler fält än det andra. Hen fick då hantera detta genom att kombinera de två dataseten genom analyskodning för att förena dem.

Vidare beskriver R3 och R4 att det kan vara svårt att arbeta med stora datamängder som inte innehåller samma variabler, då detta kräver en hel del utveckling. R3 berättar att hen och R4 håller på att bygga nya modeller som är mer omfattande och generella för att de ska kunna klara av flera olika cases, designs och situationer. Dessa modeller kan då till exempel ta bort variabler om datan innehåller onödig och överflödig information. R4 betonar betydelsen av att hantering av utmaningen med Data Consistency måste ske organisatoriskt. Hen menar att organisationen till en början måste sätta en standard internt på hur man vill att datan ska se ut. Därefter är det viktigt att kontrollera att datasetet följer den standarden i stabiliseringsprocessen som man har beslutat om. Skulle datasetet inte vara enligt standarden kan man använda olika visualiseringsverktyg, exempelvis som Google Cloud för att formatera om datan.

Både R1 och R2 anser att Data Consistency inte är av ett större problem inom deras organisationer. R2 berättar att på F2 följer de mallar som använts genomgående under många år. Däremot beskriver R2 att de har haft problem med att koppla olika faser till varandra inom samma projekt, exempelvis försäljnings- och implementationsfasen. R2 förklarar att så fort F2 var förbi försäljningsfasen så “släppte” man det och gick vidare till nästa fas utan att koppla samman datan. Detta resulterar i att datan inte är exakt likvärdig, trots att det fortfarande är inom samma projekt. Företaget hanterade utmaningen genom att låta en person sitta manuellt och försöka koppla ihop dem. Till sist fick F2 strunta i kopplingen då det ansågs vara för svårt. R2 poängterar däremot att teamet informerade F2 om problematiken vilket resulterade i att företaget i framtiden skulle försöka skapa denna koppling.

Vidare beskriver R1 att dimensionen inte är ett problem för dem till följd av att de har lång erfarenhet av traditionella analyser såsom Business Intelligence. Vid de traditionella analyserna lägger man stor kraft vid inladdning av data och använder sig exempelvis av metoder såsom

Constraints och Single Point of Truth. Däremot menar R1 att hen och hens team till viss del är emot denna typ av filtrering, då de anser att intressanta detaljer av datan kan försvinna.

4.6 ETL

Varken R2, R6 eller R7 visste vad ETL var för något och svarade därför att de inte visste huruvida det var en process som används på företagen de arbetar på. R5 däremot beskrev ett exempel av en ETL-process och berättade att hen hade erfarenhet av det. Hen menar att ETL-processen gör det möjligt för dem att testa all data från en data population, vilket annars kan vara en utmaning. Hen förklarar att det även hjälper till att öka kvaliteten eftersom de då kan hävda att de analyserat, profilerat och testat 100% av all data. Vidare menar R5 att transformationen av data underlättar om de exempelvis får rådata i ett dåligt strukturerat format. Då hjälper transformationen att göra datasetet mer läsbart även för dem som inte är dataspecialister.

R1 förklarar att deras ETL-process har en ganska strikt Constraint-hantering i två steg mellan dotterbolagen och huvudbolaget som hjälper till att transformera och filtrera all data. Detta bidrar i förlängningen till bättre datakvalitet. Hen menar dock att det finns risk att vissa intressanta data filtreras bort i processen, vilket gör att man förlorar information och detaljer på vägen.

Enligt R3 har ETL varit ”ett hett begrepp” (Bilaga 3, Rad 30) inom Data Management. Vidare beskriver R3 att hen och R4 har kollegor som arbetar mycket med det. R3 berättar även att de har använt sig av ETL i mindre skala, men att det inte är något som normalt ingår i deras arbetsområde. De förklarar att de använt olika ETL-flöden för att få fram det format de behöver till en definierad datamodell. Med andra ord har det använts för att hantera utmaningen med att utvinna, transformera och lagra datan på ett lämpligt sätt.

4.7 Urval av Data Samples

Utmaningen med urval av Data Samples var något de flesta av respondenterna inte tyckte var en svår utmaning att handskas med. För R2, som oftast inte jobbar med lika stora datamängder som resten av respondenterna, delas datan endast upp i samples om träningsdata och testdata. Hen beskriver dock att modellerna kan skifta ganska mycket beroende på vilket sample som väljs eftersom det är så pass små datasets. Detta hanteras genom att använda samplet ändå och försöka få ut ett medelvärde att basera prediktionen på.

R5 tycker att en utmaning inom samples kan vara att urvalen drivs av huruvida datan är intressant i fråga om exempelvis avvikande värden, potentiella fel eller outliers. Det leder till att urvalet av samples blir mindre fokuserat på att ta ut random samples och mer fokuserat på intressant data. R6 håller med, men påpekar även att det är väldigt sällsynt att de gör den här typen av urval av samples. När detta väl sker så används statistiska kriterier för att samplet ska bli så representativt för hela populationen som möjligt. Det är svårt att säkerställa att dessa statistiska kriterier är tillförlitliga, vilket också är en anledning till att detta görs så sällan. R5 intygar även att det vanligaste fallet i deras organisation är att använda sig av hela datasetet, för annars kan de inte garantera att datan är komplett. Vidare förklarar hen att det kan ta väldigt lång tid att behandla en så stor mängd data, speciellt att hämta all data. Men då används verktyg som SQL, Qlik Sense och Tableau som kan behandla stora volymer data.

R7 som arbetar inom revision anser också att urval av Data Samples är en utmaning och hen syftar då på när de använder sig av stickprov vid granskning av stora mängder data. Metoden som används vid samplet ska vara regelrätt och det ska kunna hålla i domstol att de har kollat på hela datat med de här stickproven. För att hantera detta och för att samples ska representera hela datamängden används olika system som plockar ut de mest väsentliga stickproven.

R1 förklarar att sampling är mycket vanligt förekommande i hens organisation, till skillnad från R2, R5 och R6. R1 anser inte att urval och samples är någon direkt utmaning utan något de gör när det blir otympligt att hantera jättevolymer av data. För att få ut ett representativt sample beskriver hen:

> … jag har ju duktiga statistiker i mitt team så dem kan säkerställa att vi kan ta ut en liksom ett 10%-sample från den stora, som är representativt med helheten och att vi kan lita på det. De resultaten vi når med de här samplet kan vi förvänta oss att vi kommer nå när vi skalar upp (Bilaga 2, Rad 44).

Efter följdfrågor klargörs att hen inte i detalj vet hur de hanterar detta men att det används statistiska modeller som finns exempelvis i Python-bibliotek och som konfigurerar sig utifrån storlek och typ på datasetet. Hen betonar dock att det är viktigt att någon med rätt kompetens genomför detta moment då det lätt kan bli fel och vilseledande, vilket kan ge ett skevt sample.

R4 anser inte heller att urval av Data Samples är en särskilt stor utmaning då de ofta använder sig av någon av molntjänsterna som Google eller Microsoft Azure där det finns inbyggd funktionalitet som genomför statistisk sampling av datan. Hen förklarar att det är möjligt att välja mellan olika typer av samplingsmodeller som exempelvis random sampling, att välja en variabel som riktlinje eller att ange en viss fördelning. Det är mer en drag and drop-process som inte brukar skapa några svårigheter. R3 intygar detta och kan inte komma på att hen hittills har stött på att modellerna varierar mycket beroende på vilket sample som används.

4.8 Felmarginaler

Samtliga respondenter är överens om att det kan vara acceptabelt att det förekommer en viss felmarginal i den data som används, så länge den är tydligt identifierad. R2 och R5 understryker att det är väldigt ovanligt att ett dataset inte innehåller några fel överhuvudtaget. Respondenterna använder dock olika metoder för att identifiera felmarginaler i datan. Både R2, R3 och R4 beskriver att de använder sig av olika visualiseringar av datan för att identifiera felmarginaler. R2 förklarar att de brukar börja med att ta fram ett histogram som de använder sig av för att sedan identifiera ett tröskelvärde som fungerar som en gräns för hur stor felmarginal som tillåts.

Vidare berättar R4 att de använder sig av Google- och Microsoftverktyg som kan ge dem visualiseringar och statistik på felmarginaler i deras data. Dessa verktyg kan tillhandahålla statistik på hur fullständig deras data är, till vilken grad den följer korrekt format och hur många ”records” som saknas. Utöver detta ger verktygen även exempel på vilka åtgärder som kan tas för att förbättra datakvaliteten.

Både R5, R6 och R7 beskriver att de i likhet med R2 använder sig av tröskelvärden i form av väsentlighetstal för att bestämma hur stor felmarginal som tillåts i deras data. R5 förklarar att de använder sig av CCT (Clearial Trival Threshold) för att kontrollera om datan håller sig inom ramen för vad som anses vara tillåtet. Termen CCT är enligt R6 ett begrepp som används specifikt inom revision och hen förklarar det på följande vis:

> It's a numerical value that we state which we find is a tolerable misstatement for a company's financial reports without them being so misstated that they do not represent the economical standpoint of the company (Bilaga 5, Rad 9)

Enligt R1 och R4 kan felaktig data eventuellt rengöras om det finns möjlighet till att förbättra datakvaliteten på så vis. I likhet med detta uttrycker R3, R4 och R7 att de i första hand försöker korrigera felen om möjligt. R6 berättar att de kan använda sig av dashboards för att undersöka fel i datan och ta reda på vad eventuella avvikelser beror på. Om det går att hitta en rimlig förklaring till avvikelserna kan de sedan välja om de vill utesluta datan från analysen eller acceptera dem.

Den vanligaste åtgärden är helt enkelt att ta bort den data som är felaktig och utesluta den från analysen enligt R2, även om det inte nödvändigtvis alltid är den bästa åtgärden. I motsats till detta berättar R7 att det är väldigt ovanligt att de tar bort data då hen arbetar med revision. Vidare förklarar hen att till skillnad från andra organisationer så tar de inte bort avvikelser som Outliers i sin data inom revision, utan de ser istället dessa som extra intressant data som behöver undersökas mer i detalj.

Slutligen beskriver både R1 och R2 vikten av att presentera eventuella felmarginaler som förekommer i den data som använts när resultat från en analys ska tolkas. Detta presenteras i syfte att hjälpa dem som ska tolka analysresultatet att få en bättre förståelse för och mer rättvis bild av utfallet.

4.9 Datarengöring och korrekt representation av data

Alla respondenterna ansåg att datarengöring och korrekt representation av data var en utmaning, men hanterade detta på olika sätt. R1 menar att det är svårt att förstå vad som egentligen är smutsig data. Hen menar att det är svårt för dem att få en full förståelse för detta då de inte arbetar operativt med de delarna i verksamheten. Således anser R1 att business-förståelsen är en viktig del i arbetet, vilket R1 och hens team hanterar genom att fråga en anställd på just den delen av verksamheten. R7 är inne på samma spår som föregående respondent, att en svårighet med datarengöring är att organisationen missar observationer och data som är väsentliga för modellen samt det ändamål organisationen vill komma fram till. Vidare menar hen att det inte läggs mycket tid på att utvärdera och undersöka om datan som tas bort vid rengöringen var korrekt eller inte.

För att hantera utmaningen med data rengöring och korrekt representation av data beskriver R5 att de aldrig ändrar på den underliggande råa datan, utan ändrar endast kosmetiskt så att det blir lättare att bearbeta. Det kan till exempel vara att lägga till kolumner eller addera andra saker, men aldrig ändra det underliggande innehållet. Om det till exempel inte går att garantera att den rengjorda datumkolumnen blir helt korrekt, så kommer de inte heller basera analysen på den kolumnen. Hen betonar även vikten av att någon som kan datan kontrollerar så att den råa datan inte av misstag har manipulerats under rengöringen. I likhet med R5 beskriver R2 och hens team inte heller ändrar för mycket i rådatan utan att försöka behålla den som den ser ut, då det till och med finns skrivskydd på datan.

Andra sätt att kontrollera så att datan är komplett efter rengöringen menar R6 är att kontrollera att den fortfarande innehåller samma antal rader och kolumner, har önskat innehåll samt är i rätt format. Det kan göras genom script som räknar och jämför rader eller genom att sortera datan i tabeller och jämföra så att innehållet är rätt. Dock menar hen att det inte är möjligt att kontrollera varje enskild rad vid stora mängder data, vilket beskrivs nedan:

> … you can never have 100% assurance or confidence that you haven't altered anything if you have a very large dataset say 10 million rows, you can't check every row. At some point you can satisfy that it is accurate enough after it being transformed (Bilaga 6, Rad 79).

Ett konkret exempel på hur R1 och hens team säkerhetsställer att de får en korrekt representation av datan är att de inte lägger det förväntade storleksintervallet på siffrorna alltför snävt och därmed raderar för många rader. Om storleksintervallet är alltför snävt kan detta resultera i att man exkluderar ett potentiellt kundsegment. Ett annat exempel som R1 berättar om för att garantera korrekt representation av datan är ifyllning av tomma fält. Hen understryker vikten av att den som fyller i de tomma fälten besitter rätt kompetens och använder sig av statistiska tekniker i och med att man konstruerar data i efterhand. R7 beskriver att de säkerhetsställer detta i den prediktiva modelleringen genom att jämföra ett träningsdataset med ett testdataset för att garantera att träningsdatan representerar testdatan på ett korrekt sätt.

Vidare beskriver R4 att de tacklar problemet genom att definiera en standard för vilket format som krävs för uppdraget eller byggnaden utav modellen, innan hen utför datarengöringen. Hen berättar att "så länge man förhåller sig till den standarden som är satt från början så kan man säkerställa att datan håller den som du uttryckte det att den fortfarande representerar" (Bilaga 4, rad 67).

Alla respondenter var eniga om att kostnaden associerad med datarengöring inte var något som organisationerna sparade in på då detta är en viktig och avgörande del för slutresultatet av Big Data Analytics. R1 berättar att det innebär stora risker att inte rengöra datan i och med att fel beslut kan tas baserat på analyser, vilket i slutändan resulterar i ännu större kostnader. Respondenterna från konsultföretagen beskriver att den stora kostnaden för datarengöringen är i form av tid, särskilt inom deras bransch då det handlar om betalning per timme. R3 menar därför på att de flesta kunderna oftast redan ha genomfört rengöringen av data själva, innan konsulter kommer in och utför avancerade analyser. R5 menar även att datarengöring alltid är en del av analysprocessen och att antal timmar och resurser är inkluderat i budgeten för projektet så att det finns utrymme att rengöra datan.

4.10 Ledningens stöd och engagemang

Huvudparten av respondenterna tycker att de får stöd och engagemang utav ledningen i att upprätthålla en god datakvalitet inom Big Data Analytics. R5, R6 och R7 som jobbar på konsultföretag utgick från ledningen hos deras kunder som de har arbetat för och understryker därmed att stödet och engagemanget varierar från kund till kund. R5 beskriver följande:

> …I think it really depends on the actual client because sometimes the dataset is very or the system is very old for example if maybe it's a smaller company they just don't have the time and resources to fix the data or investing in a completely new data extern in order to fix the data (Bilaga 5, rad 84).

Vidare menar R5 att ledningen ibland endast är intresserade att få datan korrekt för att de är ansvariga inom dessa områden, exempelvis på grund av GDPR. I tillägg till detta menar R7 att ledningens stöd hos kunden mer beror på företagets mognad och kompetens internt.

En svårighet inom frågan kan vara att erhålla och utvinna data då det förekommer motstånd från ledningen om de inte förstår vad analytikerna vill åstadkomma menar R5. En annan utmaning är att de från IT-avdelningen som utvinner datan inte förstår varför man behöver ett specifikt dataset. Respondenten menar då att man bemöter detta genom att ha en konversation tillsammans innan analysen, för att få en gemensam förståelse. Vidare menar R5 att ledningen är stöttande i analysarbetet ifall de finner något i resultatet intressant och insiktsfullt. Hen menar att de genom visualiseringar och rapporter måste övertyga ledningen om att de kommer få nytta av resultatet.

Fortsättningsvis refererar R4 till att verksamheter som tillämpar Big Data Analytics ofta har en organisatorisk struktur som stöttar data- och informationshantering inom företaget. Hen menar således att organisationer som begär hjälp med sådana tekniker ofta har kommit långt i sin digitaliseringsresa. Vidare understryker R4 vikten av att sätta en organisationsstruktur där data anses ha stort värde och vara en resurs inom företaget samt ha en person som är ansvarig för datahanteringen.

R1 delar uppfattningen om att de får stöd och engagemang av ledningen på F1, då företaget satsar och investerar mycket på strategisk nivå inom Big Data, Data Science och AI. De får stöd i form av budgetar och anställning av personer med rätt kompetens. Däremot betonar respondenten på att de inte får stöd specifikt till just datakvalitet utan mer till själva analysen, där datakvaliteten ligger som grund. Med andra ord får de indirekt stöd till datakvaliteten då den är en del utav Big Data Analytics.

Till skillnad från resterande respondenter anser R2 att ledningen på F2 lägger mer fokus på att få ut mycket resultat, snarare än att få resultat utav hög kvalitet inom Big Data Analytics. R2 uttrycker det som att man tillämpar synsättet kvantitet före kvalitet.

5 Diskussion

I följande avsnitt diskuteras och analyseras studiens litteraturgenomgång i förhållande till det empiriska resultatet som tagits fram utifrån de kvalitativa intervjuerna. Detta avsnitt syftar till att analysera hur de utmaningar som har identifierats och presenterats i tidigare avsnitt hanteras hos organisationer i praktiken.

5.1 Information Completeness

Det framgår inget specifikt kring hur utmaningar med Information Completeness bör hanteras i litteraturen och därav är det inte heller möjligt för oss att göra en jämförelse mellan teori och praktik i denna aspekt. Vidare visar inte resultatet av intervjuerna på att det skulle finnas någon allmän hantering av utmaningar inom dimensionen som samtliga respondenter använder sig av. Detta kan grunda sig i att respondenterna inte angav samma anledningar till utmaningarnas uppkomst, vilket skulle kunna bero på att de är verksamma inom skilda branscher. Deras olika bakgrunder medför troligen att de har skilda erfarenheter och därmed även åtskilda perspektiv.

I likhet med vad litteraturen indikerar uppfattar respondenterna att utmaningar inom denna dimension förekommer i deras arbete. Trots detta menade tre av respondenterna att detta inte var någon stor utmaning för dem. Detta kan indikera på att de antingen har ett välfungerande arbetssätt för att hantera denna typ av utmaning, alternativt att de arbetar i en bransch som är mindre utsatt för förekomst av ofullständiga data.

Två av respondenterna som arbetar i revisionsbranschen menade att de i normalfallet inte upplevde problemet med saknade värden, med anledning av att det finns branschspecifika regleringar inom revision som kräver att alla uppgifter är fullständiga. Det finns därför anledning att tro att det inte är branschens resistens mot ofullständiga data, utan tvärtemot dess sensitivitet för det som gjort att deras organisation hanterar denna utmaning på ett sådant sätt att det inte uppfattas som ett problem för respondenterna. Resterande respondenter inom revisionsbranschen uttryckte dock inte samma uppfattning, utan pekade på extraheringen av data som en problemfaktor för dimensionen. De beskrev att de i första hand lät sina kunder sköta dataextraheringen och först om eventuella fel uppstått erbjöd de sig att hjälpa kunden göra om processen. En anledning till att de två förstnämnda respondenterna inte beskrev detta som en problemfaktor skulle kunna innebära att de har en bättre hantering av denna process.

Samma två respondenter diskuterade dock att det är nödvändigt att göra en avvägning i hur kostsam avsaknad av viss data är och att det kan vara acceptabelt att inte alltid ha all data tillgänglig. Anledningen till att svaren skiljde sig åt beror troligtvis på att de här syftar på andra delar av sitt arbete som inte omfattas av tidigare nämnd reglering.

Tre av respondenterna använde sig av modeller för att hantera utmaningen med Information Completeness. De använde dock inte modeller i samma syfte då en av dem använde det för att hantera bristen på tillräckligt detaljerad data med hjälp av Ensemble modell, vilket är en

utmaning som diskuteras av Pipino, Lee och Wang (2002). De två andra respondenterna använde istället modeller i syfte att identifiera saknade värden i sin data, vilket är en vanligt förekommande utmaning inom Information Completeness enligt Sidi et al. (2012). Det skilda arbetssättet med modeller grundas troligen i att deras förutsättningar sett väldigt annorlunda ut, beroende på den data de har haft tillgänglig.

5.2 Data Accuracy

Data Accuracy innebär som Sidi et al. (2012) beskriver den omfattning som datan är korrekt, pålitlig och certifierad. Detta var något alla respondenter kände igen sig i och ser som en utmaning i deras arbete. För majoriteten av respondenterna sätts de ofta i ett val där de ska fatta beslut om de ska korrigera de fel som finns, eller om de ska lämna datan som den är. Det kan till exempel vara att lämna tomma värden blanka. Om de väljer att korrigera de fel som finns så hanteras det genom att ta kontakt med den person som har fört in datan manuellt, visualisera datan på en dashboard, använda sig av kolumner med Previous Values så att man kan återskapa den förra raden samt hitta ledtrådar och slutligen att göra en prognos och förklaringar av datan.

Om man istället väljer att låta datan vara så hanteras detta genom att exkludera den delen av data i analysen samt att tydligt beskriva det i presentationen av analysen. Man kan också välja att ta med datan i analysen i alla fall men då även tydligt dokumentera att det kan finnas felaktigheter i datan. Detta kopplar vi samman med utmaningen med felmarginaler, som handlar om att man ibland får acceptera att det finns en viss felmarginal då det är svårt att få datakvaliteten helt perfekt (Baesens et al., 2016). I både detta fall och utmaningen med Data Accuracy beskriver respondenter att det är helt okej att ha felmarginaler i datan, vilket även litteraturen beskriver (Baesens et al., 2016). Det som vissa respondenter dock ansåg är väldigt viktigt är att dokumentera och tydligt framföra att dessa fel kan finnas i datan så att den som tolkar analysen får en sanningsenlig bild. Antingen då förklara varför man exkluderat en del av datan i analysen eller förklara vilka delar som kan vara fel samt vilka fel. Detta är något som litteraturen inte nämner specifikt för Data Accuracy, men som betonas under felmarginaler av Baesens et al. (2016).

En annan aspekt som respondenterna nämner är svårt inom Data Accuracy är just att upptäcka felen inom detta. När man arbetar med så stora mängder data går det inte att kontrollera varje rad, och det är då nästan omöjligt att identifiera fel som fortfarande är i samma format. Detta är inte heller något vi anser att litteraturen nämner, trots att det verkar vara vanligt förekommande och svårt att lösa. Vi anser att ett sätt att förebygga detta kan vara att kontrollera data vid inmatningen eller extraktionen av datan så att färre av dessa fel uppkommer.

5.3 Data Currency

Utifrån resultatet går det att dra slutsatsen att respondenterna inte anser att Data Currency är en större utmaning. Således skiljer sig teorin mot praktiken i denna dimension. Alla respondenter menade att de använder sig av up-to-date data. Två respondenter menade att utmaningen hanteras genom kontroller på rådatan. Ett annat sätt en respondent hanterade utmaningen med Data Currency var att inte använda sig av datan om hen ansåg att den var för gammal och orelevant. Detta kan utläsas i empirin genom exemplet när respondenten valde ut samples för att träna en modell i ett specifikt projekt. Dilemmat var då hur gamla de tidigare projekten kunde vara för att

kunna användas som samples för datan. Här behövde respondenten göra en avvägning kring hur gammal data de skulle använda sig av. Risken med att använda sig av gammal data var att den skulle kunna vara irrelevant, men fördelen med detta var att de fick mer data att arbeta med. Alternativt kunde de bestämma sig för att inte använda sig av data från de äldre projekten och därmed vara begränsade till en mindre mängd data. Då respondenten beslutade sig för det senare alternativet går det emot litteraturen där Sidi et al. (2012) menar att data kan vara aktuell och relevant trots eventuella avvikelser orsakade av tidsrelaterade förändringar. Respondenten menar att det har skett för stora förändringar de senaste tio åren vilket medför att den äldre datan skulle vara irrelevant.

I litteraturen menar Yu (2013) på att tidsstämplar ofta är felaktiga eller otillgängliga och att det därför är en stor utmaning inom Data Currency. Korrekta tidsstämplar är därmed ett sätt att hantera denna utmaning på, vilket en av respondenterna som verkar inom detaljhandeln nämner att de använder. I teorin beskriver Yu (2013) att sammanfallande datumvärden kan vara ett problem i praktiken. Respondenten menar att deras tidsstämplar är överensstämmande i och med att de sällan används till en högre precision än en timma. Även en annan respondent hanterade problemet med Data Currency med hjälp av tidsstämplar. Då detta gör det möjligt att urskilja vilken data som är mest aktuell genom att jämföra tidsstämplarna på datan och på så sätt kontrollera dess giltighetstid.

5.4 Data Deduplication

Dimensionen Data Deduplication beskrivs länge ha varit en utmaning i litteraturen och kan tänkas vara den datakvalitetsdimension som studeras och undersökts mest (Fan, 2015). Efter att ha intervjuat våra sju respondenter så verkar just denna utmaning inte vara särskilt vanligt förekommande, samt att det var ett lätthanterat problem. Fyra av sju respondenter känner till utmaningen men anser inte att den är vanlig i deras arbete. Hur denna utmaning hanteras i praktiken skiljde sig beroende på organisation. Exempel på sätt att hantera dubbletter i datan är genom Constraint-kontroller i databasen som fångar upp dubbletterna, använda SELECT DISTINCT vid queries till databasen, gruppera dubbletter till en rad samt att föra register på vad datan innehåller och sedan göra en jämförelse. En anledning till att litteraturen säger att detta är den vanligaste och svåraste utmaningen inom datakvalitet hade kunnat vara för att det länge har varit en svår utmaning som organisationer kämpat med, men sedan Fan (2015) skrev detta så har det utvecklats lösningar på det. Exempelvis kan det ha tagits fram verktyg och metoder som organisationer idag använder, och därför anser inte heller våra respondenter att detta är en svår utmaning att hantera.

Fan (2015) nämner även att det är nödvändigt och viktigt att korrekt identifiera rader från olika källor för att sedan förena datan och förbättra informationen. Vidare beskriver han också att konflikter ofta uppstår under integreringar av olika datakällor. Detta är något majoriteten av respondenterna även själva nämner, alltså att dubbletter främst uppstår när man för samman data från olika system. Hur detta hanteras är genom manuell handpåläggning och att ha en unik identifierare av varje rad, vilket också är det Fan (2015) beskriver. Respondenter berättar även att det finns verktyg som hanterar detta och ett exempel som nämns är Google Clouds plattform.

5.5 Data Consistency

Inom Data Consistency är litteraturen och empirin överensstämmande, då majoriteten av respondenterna menade på att dimensionen var en svårighet för dem. En iakttagelse utifrån empirin är att alla respondenterna som identifierade dimensionen som en utmaning jobbade på konsultföretag medan de respondenterna som inte ansåg det som en utmaning jobbade inom industri- och dagligvaruhandelföretag. Detta skulle kunna motiveras genom att konsultföretagen oftast får sin data externt från andra organisationer med olika typer av system, vilket innebär att datan kan komma i fel format.
En respondent beskrev att de bemötte utmaningen med att ge rekommendationer till kunden angående vilket format de önskade att få datasetet i. Om respondenten ändå erhöll data i fel format hanterades detta genom att tvätta- och transformera datan till det önskade formatet. Denna typ av hantering går att likställa med vad som görs i ETL-processen. Souibgui et al. (2019) menar på att ETL-processen är en process som hanterar data med olika format och kvalitetssvårigheter, där T:et står för transform som innefattar konvertering av data till korrekt format. Vidare menar respondenten att efter transformationen så sker kontroller för att se till att datan fortsatt har samma betydelse som innan omformateringen. Detta kan kopplas till litteraturen där Clarke och Taylor (2018) understryker vikten av att se till att datan behåller samma innebörd efter datarengöringen.

Två av respondenterna bemöter utmaningen med Data Consistency genom att själva utveckla modeller som är mer generella inom sitt område och klarar av data med många olika format. Modellerna kan därmed klara av många olika situationer utan att de uppstår svårigheter när de erhåller data i fel format. I en annan situation upptäckte en av respondenterna att hen erhållit två dataset av samma data, men som var av olika format. Detta går i linje med vad Fan (2015) menar att innebörden av Data Consistency är, att upptäcka fel och inkonsekvenser i datan. Situationen bemöttes genom kodning av analysen i syfte att förena dataseten eller med metoder såsom Constraints och Single Point of Truth.

Resultatet av empirin säger att utmaningen med Data Consistency måste hanteras organisatoriskt, genom att bestämma en intern struktur på datan och sedan kontrollera att detta efterföljs. Detta skulle kunna kopplas till litteraturen där Baesens et al. (2016) beskriver betydelsen av att ledningen förstår tyngden av hög kvalitet på data. Taskin et al. (2019) understryker även vikten av en datadriven kultur. Även Fass (2018) skriver att finansledningar kämpar hårt med att hantera datakvaliteten inför analyser. Utifrån detta kan man således anta att utmaningen med Data Consistency måste tas i beaktning och hanteras på en högre nivå av ledningen och inte bara av de som utför analyserna. När strukturen finns på plats skulle det kunna innebära att svårigheter med dimensionen på en operativ nivå i organisationen minskar.

5.6 ETL

Enligt litteraturen är ETL-processen nödvändig för att inhämta data av hög kvalitet och för att i förlängningen kunna genomföra en noggrann relevant analys (Souibgui et al., 2019). Trots detta hade tre av respondenterna aldrig hört om begreppet ETL tidigare och var därför ovetandes om av vad det innebar. Detta skulle antingen kunna bero att respondenterna inte är delaktiga i processen eller att företaget inte använder sig av ETL. Ett annat förslag är att ETL som begrepp inte

används men att själva processen med extract, transform och load fortfarande tillämpas. Då samma tre respondenter diskuterade processen för datatransformation i form av datarengöring och två av dem även diskuterade utmaningar med extrahering av data, tyder detta på att de skulle ha ett etablerat arbetssätt för att hantera denna typ av processer.

Två av de övriga respondenterna kände till ETL-processen men hade inte mycket personlig erfarenhet av den, då den främst hanterades av deras Data Management avdelning. Anledningen till att majoriteten av respondenterna inte hade någon utbredd erfarenhet av ETL, tror vi beror på det faktum att vi valt att intervjua personer i roller såsom Data Scientist och Data Analyst med erfarenhet av Big Data Analytics och inte någon som arbetar direkt med datakvalitet. Vi har därför anledning att tro att de tre steg som utgör ETL-processen i regel används, men att de i huvudsak inte tillämpas direkt av respondenterna själva, samt att termen ETL inte nödvändigtvis är ett etablerat begrepp som används.

Enligt resultatet från de två respondenter som faktiskt var familjära med begreppet används ETL i syfte att strukturera data och förbättra dess kvalitet. Detta överensstämmer väl med vad litteraturen beskriver att ETL-processen kan hjälpa till att bidra med (Souibgui et al., 2019).

5.7 Urval av Data Samples

Fisher et al. (2012) anser att även fast organisationer har stor kvantitet av data så innebär det inte att samplet är representativt för hela populationen. Inte heller att samplet ger en fullständig sanning av användarens beteende eller motiv. Litteraturen nämner dock inte hur man går tillväga för att hantera detta eller vilka metoder som används för att få ett så representativt sample som möjligt. Våra respondenter beskriver hur detta hanteras inom deras organisationer, de hade ett likt tillvägagångssätt med endast mindre skillnader. Den gemensamma metoden är att statistiska beräkningar görs på datan för att få ut ett representativt sample. För att genomföra detta används olika system och verktyg som exempelvis Python-bibliotek, Google Cloud och Microsoft Azure. Huruvida denna utmaning ansågs svår att hantera eller inte skiljde sig mycket mellan respondenterna där majoriteten menade att det inte är något problem. Några respondenter ansåg dock att det är svårt att få ut ett representativt sample oavsett och behandlar därför hellre hela datasetet, trots att detta kan ta tid och kraft.

Detta går i linje med vad Clarke och Taylor (2018) beskriver, då de hellre behandlar hela mängden data istället för att välja ut samples. De menar att tillgången till hela datasetet överträffar möjligheten till att använda sample, då det finns möjlighet att hitta nya korrelationer i sådana volymer. Det vill säga att kvantitet är viktigare än kvalitet. Respondenterna håller dock inte med om det, då vi uppfattar att de väljer att använda sig av hela datasetet just för att förbättra datakvaliteten. Syftet är snarare att datan ska ge rätt representation av verkligheten än att nya korrelationer ska identifieras. Med andra ord att de avstår från att använda samples då de inte vill att datakvaliteten ska försämras. I tillägg till detta menar Fisher et al. (2012) att oavsett hur stort datasetet är så krävs det övervägning och tolkning för att få tillräcklig kvalitet på datan. En slutsats vi har dragit i samtliga fall är att det skiljer sig märkbart hur urval av samples hanteras och om det anses vara utmanande eller inte, både i litteraturen och i praktiken.

5.8 Felmarginaler

Respondenternas uppfattning kring förekomsten av felmarginaler i deras data stämmer väl överens med litteraturens beskrivning som menar att data aldrig kommer vara av perfekt kvalitet och att det alltid kommer förekomma vissa felmarginaler (Baesens et al., 2016). Samtliga respondenter menade att det var acceptabelt att det kan förekomma fel i den data som används men att det var viktigt att tydligt identifiera en tillåten felmarginal, i likhet med vad som uttrycks i litteraturen. De använde olika benämningar men genomgående i allas svar var att de använde någon form av tröskelvärde för att hantera felmarginaler. Tröskelvärdet skulle kunna liknas med en gräns för den lägsta tillåtna kvaliteten på data som kan användas vid en analys.

Vidare beskriver litteraturen att det kan vara väldigt komplext att mäta datakvalitet då kvaliteten på data kan utvärderas olika beroende på vad den ska användas till och hur strikta krav som ställs (Baesens et al., 2016). Detta speglar resultatet från majoriteten av respondenterna väl som uttryckte att de inte var medvetna om att det ens fanns någon officiell mätning av datakvalitet på deras företag. Med undantag för identifiering av felmarginaler inför enskilda projekt eller analyser, där datakvaliteten kan sägas mätas utifrån den acceptabla felmarginalen. Respondenterna indikerade i likhet med litteraturen att kraven som ställs på datakvalitet varierar beroende på vilken sorts analys eller projekt de arbetar i och vad syftet med dessa är.

Vid tolkning av resultatet från en analys betonar litteraturen (Baesens et al., 2016) vikten av att presentera eventuella felmarginaler i den data som använts, så dessa kan tas i beaktning och bidra till en rättvis bild av analysresultatet. Det var dock endast två av respondenterna som uttryckligen beskrev att detta gjordes i samband med tolkning av resultatet. Vi tror däremot inte att det går att utesluta att övriga respondenter inte skulle ta hänsyn till felmarginaler för att bättre förstå utfallet av en analys. Detta då övriga respondenter beskrev att de var noggranna med att friskriva sig från fel i datan och att presentera felmarginaler. De talade ur ett mer allmänt perspektiv och det finns därför en möjlighet att även dem tar felmarginaler i beaktning vid tolkning av analysresultat.

5.9 Datarengöring och korrekt representation av data

Litteraturgenomgången beskriver att datarengöring innebär de tekniker som används för att uppnå hög kvalitet på datan. Clarke och Taylor (2018) anser dock att få av de processer som appliceras i praktiken involverar jämförelser med en auktoritativ extern referens och att de flesta processer bara är manipulationer som är baserat på statistiska analyser av själva datan. Efter att ha intervjuat respondenterna visade det sig att detta inte alls stämde i praktiken. Samtliga respondenter nämnde och poängterade vikten av att jämföra den rengjorda datan med den ursprungliga datan så att den har en korrekt representation. Detta görs genom att en verksamhetsexpert eller en expert på datan kontrollerar om den råa datan har manipulerats och om datan fortfarande innehåller samma antal rader och kolumner. Det säkerställs även genom att den underliggande strukturen aldrig ändras och att det definieras en standard tillsammans med verksamheten redan innan rengöringen påbörjas.

Vi tror att det som Clarke och Taylor (2018) menar kan vara en risk för många organisationer vid datarengöring. Dock verkar det som att de flesta organisationer är medvetna om denna risk och därför lägger ner mycket resurser just på att jämföra den rengjorda datan med den auktoritativa

referensen. För att kunna tackla denna utmaning anser vi att det inte endast behövs kompetens inom datarengöring och Big Data Analytics, utan det är även viktigt att någon som har förståelse för verksamheten och dess processer är involverad i rengöringsprocessen.

Beebe och Walz (2005) påpekar att datarengöring för stora mängder data har betydande och höga kostnader förknippat med det, både de operativa kostnaderna och de strategiska kostnaderna (Lucas, Ishfaq & Raja, 2014). Respondenterna var dock eniga om att kostnaden med datarengöring inte är något som det sparas in på, då det har en avgörande roll för slutresultatet. De menar att om dessa kostnader på rengöring av datan inte läggs innan själva analysen, kan det resultera i ännu större kostnader senare. Medan Lucas, Ishfaq och Raja (2014) skiljer på operativa kostnader och strategiska kostnader så pratar våra respondenter mer om kostnader i form av tid och konsulttimmar. De menar att det snarare blir en fråga om kunden vill lägga stora kostnader på att ta in konsulter som genomför datarengöringen, eller om de i bästa fall har tillräcklig kompetens för att genomföra datarengöringen själva innan konsulter kommer in och utför avancerade analyser. Därmed drar vi slutsatsen att besluten inte baseras på huruvida organisationen ska lägga kostnader på datarengöringen eller inte, utan snarare om det ska ske internt inom organisationen eller externt genom konsulter.

Kostnaderna med datarengöring verkar därför inte vara en utmaning som organisationer behöver hantera utan ses snarare som en självklar del av analysprocessen och därmed inkluderas i budgeten. Lucas, Ishfaq och Raja (2014) skriver dock att utmaningen är att balansera kraven för datakvaliteten och resursbegränsningarna när det gäller tid, kostnad och expertis vilket är precis det respondenterna menar.

5.10 Ledningens stöd och engagemang

Alla respondenter förutom en ansåg att de fick det stöd de behövde av ledningen inom datakvalitet för att genomföra sina uppgifter. En respondent menade att Big Data Analytics är ett stort satsningsområde på strategisk nivå för dennes organisation. Detta ligger i linje med vad som beskrivs i litteraturen där Fass (2018) menar på att Data Analytics är en av ledningens topp-prioriteringar. Däremot poängterar respondenten att det inte är specifikt datakvalitet som ledningen visar stöd och engagemang för utan mer de delar som god datakvalitet möjliggör. Detta går till viss del emot vad Baesens et al. (2016) betonar, då de menar att det är av stor betydelse att ledningen förstår vikten av hög kvalitet på datan då det är grunden i analyser.

En respondent menade att det är av stor vikt att det finns en organisatorisk struktur som stödjer informations- och datahantering för att hantera utmaningen. Vidare antyds det att organisationen måste ha en kultur där data anses vara av stort värde och resurs för företaget. Detta understryks även av Taskin et al. (2019) som menar på att det krävs att det finns en etablerad datadriven kultur för att erhålla ledningens stöd i datakvalitet inom Big Data Analytics.

Fortsättningsvis berättade en respondent att de hanterade utmaningen med att få ledningens stöd och engagemang genom att ha en konversation mellan någon från ledningen och någon från IT-avdelningen, för att få en gemensam förståelse till varför en analys ska genomföras och vad den kommer att bidra till. Även detta kan kopplas till litteraturen där Taskin et al. (2019) beskriver att öppen kommunikation är en viktig aspekt för att övervinna ledningens motstånd. Dessutom kommunicerade respondenten vad för nytta ledningen skulle erhålla från analyserna genom visualiseringar och rapport. Detta resulterade i att ledningen då visade mer stöd.

6 Slutsats

En förutsättning för att lyckas med Big Data Analytics är att den data som analyseras är av hög kvalitet. Att det finns utmaningar som behöver hanteras gällande datakvalitet inom Big Data Analytics är något som både litteraturen och resultatet av vår studie visar på. Hur dessa utmaningar hanteras skiljer sig däremot avseendevärt åt. Syftet med denna uppsats är att beskriva hur organisationer hanterar utmaningar med datakvalitet inom Big Data Analytics. De utmaningar vi identifierat i litteraturen är: Information Completeness, Data Accuracy, Data Currency, Data Deduplication, Data Consistency, urval av Data Samples, felmarginaler, datarengöring samt ledningens stöd och engagemang. I tillägg till detta har vi även kontrollerat om organisationerna använde sig av ETL i syfte att hantera utmaningarna inom datakvalitetsdimensionerna, då litteraturen säger att det är vanligt förekommande.

Information Completeness hanteras genom att (1) sammanställa nödvändig data i ett strukturerat format, (2) använda sig av Ensemble modeller för att ta fram ett generellt medelvärde om datan inte har tillräcklig bredd, (3) bestämma tröskelvärde för sina modeller för att kontrollera osäkerheter i datan, (4) använda RowCount för att kontrollera om värden saknas och (5) jämföra den extraherade data med ursprunglig data.

Data Accuracy hanteras genom att (1) visualisera data, (2) jämföra och korrigera dataset efter ursprunglig data, (3) kontrollera sin data med en expert för att korrigera felaktiga värden och hitta ledtrådar, (4) exkludera felaktig data i analysen och tydligt dokumentera detta, (5) inkludera felaktig data i analysen och tydligt dokumentera detta och (6) göra en prognos och förklaringar för datan.

Data Currency hanteras genom (1) korrekta tidsstämplar, (2) kontrollera datans giltighetstid, (3) genomföra kontroller av rådata och (4) utesluta data som inte är aktuell och relevant.

Data Deduplication hanteras genom att (1) använda Constraint-kontroller för att fånga upp dubbletter, (2) att använda Select Distinct queries eller annan kodning för att gruppera raderna till en, (3) föra register på vad dataset innehåller, (4) använda RowCount för att räkna rader med samma värden, (5) se till att varje rad i en tabell har en unik identifierare samt (6) att använda Google Cloud plattform och andra liknande verktyg.

Data Consistency hanteras genom att (1) ge rekommendationer på dataformat till kund, (2) rengöra och transformera data till korrekt format, (3) använda sig av standardiserade formatmallar, (4) använda kodning för att förena olika dataset med skilda format, (5) bygga generella modeller som accepterar data av olika format, (6) sätta en intern organisatorisk standard för dataformat, (7) använda olika visualiseringsvertyg för att formatera om data och (8) använda Constraints-kontroller och Single Point of Truth.

Urval av Data Samples hanteras genom att (1) använda statiska beräkningar för att ta ut representativa sample för hela populationen, (2) använda statistiska modeller som finns exempelvis i Python-bibliotek, Google eller Microsoft Azure för att ta ut representativa sample för hela populationen, (3) ta ut ett medelvärde om modellerna skiftar beroende på vilket sample

som används, (4) använda verktyg som SQL, Qlik Sense och Tableau för att kunna bearbeta hela data populationer vid analys istället för samples och (5) genomföra väsentliga stickprov.

Felmarginaler hanteras genom att (1) använda visualiseringar såsom histogram, (2) identifiera tröskelvärden, (3) använda Google och Microsoft-verktyg som tillhandahåller statistik på hur fullständig ens data är, (4) om möjligt försöka rengöra datan, (5) utesluta felaktig data från analysen och (6) dokumentera felmarginaler så de kan tas i beaktning vid tolkning av analysresultat.

Datarengöring och korrekt representation av data hanteras genom att (1) identifiera vilken data som behöver rengöras, (2) inkludera en person som har förståelse för verksamheten eller är experter på den data som används, (3) aldrig ändra på datans underliggande struktur, (4) använda RowCount för att kontrollera att datan innehåller samma antal rader efter rengöringen som innan, (5) sortera den rengjorda datan i tabeller och kontrollera så innehållet är korrekt, (6) inte lägga det förväntade storleksintervallet på sin datapopulation för snävt så för många rader utesluts, (7) använda sig av statistiska tekniker vid efterkonstruktion av data, (8) definiera ett standardformat innan rengöringen påbörjas och (9) se till att datarengöringen alltid är inkluderat i budgeten för hela Big Data Analytics projektet.

Ledningens stöd och engagemang hanteras genom att (1) ha en konversation med både IT-avdelningen och ledningen för att skapa en gemensam förståelse, (2) övertyga ledningen om att de kommer få nytta av analysresultat med hjälp av visualiseringar och rapporter och (3) etablera en organisatorisk struktur som stöttar data- och informationshantering.

ETL är en process som används för dataintegration och som kan hjälpa till att hantera utmaningar inom de olika datakvalitetsdimensionerna. Majoriteten av respondenterna uttryckte att organisationerna de arbetar på använder sig av ETL.

Resultatet visar på att det inte finns någon helt enhetlig hantering av respektive utmaning som alla organisationer tillämpar. Detta kan grunda sig i att respondenterna inte angav samma anledningar till utmaningarnas uppkomst, vilket skulle kunna bero på att de är verksamma inom skilda branscher och har olika yrkesroller. Det var inte heller alla respondenter som ansåg att utmaningarna var problematiska för dem, då de redan har väletablerade arbetssätt och metoder för att hantera dessa.

Bilaga 1 - Intervjuguide

Bakgrund

1. 1. I vilken roll arbetar du som?
    a. 1.Hur länge har du arbetat med det?
    b. Vad har du för arbetsuppgifter och ansvarsområden?

Olika dimensioner av datakvalitet

2. Vilka utmaningar finns det inom Information Completeness? Det vill säga i vilken utsträckning data inte saknas och har tillräcklig bredd och djup för den aktuella uppgiften.
    a. Hur hanterar ni dessa?
3. Vilka utmaningar finns det inom Data Accuracy? Det vill säga att attributvärden är exakta rent språkmässigt och har korrekt detaljnivå. Detta Accuracy innebär till den omfattning som datan är korrekt, pålitlig och certifierad.
    a. Hur hanterar ni dessa?
4. Vilka utmaningar finns det inom Data Currency? Det vill säga huruvida data är aktuell.
    a. Hur hanterar ni dessa?
5. Vilka utmaningar finns det inom Data Deduplication? Det vill säga att det bara får finnas en enda tupel med korrekta värden för en entitet och därmed ersätta dubbletter av tupler som refererar till samma riktiga entitet.
    a. Hur hanterar ni dessa?
6. Vilka utmaningar finns det inom Data Consistency? Det vill säga i vilken utsträckning data presenteras i samma format och är kompatibelt med tidigare data.
    a. Hur hanterar ni dessa?

ETL

ETL är ett vanligt begrepp inom beslutsstödsvärlden som står för Extract, Transform och Load.

7. Använder ni er utav ETL för att säkerställa datakvalitet inom Big Data Analytics?
    a. Finns det några utmaningar med ETL?
    b. Hur hanterar ni dessa?

Urval av data samples

I litteraturen står det att även fast företag har stor kvantitet av data innebär det inte att samplet de använder sig av är tillräckligt representativt för hela populationen, alltså större är inte bättre.

8. Använder ni er av data samples när ni gör analyser?
    a. Hur hanterar ni utmaningen med att välja ut data samples som ger en rättvis representation av datan?

Felmarginaler

I litteraturen poängteras vikten av att identifiera felmarginaler i ens data, då de menar att datakvaliteten aldrig kommer kunna vara perfekt.

9. Hur ställer du dig till att det kan förekomma felmarginaler i den data som används?
10. Hur hanterar ni dessa felmarginaler?
11. Mäter ni er datakvalitet och i så fall hur?

Datarengöring och korrekt representation av data

Enligt litteraturen kan datarengöring användas för att lösa problem med saknade värden, syntaktiska fel i datainnehållet, syntaktiska skillnader mellan jämförbara dataobjekt, ursprunglig låg kvalitet av data, försämrad datakvalitet under lagring och avsaknad av metadata.

12. Hur använder ni er av datarengöring?
13. Vilka utmaningar anser du att det finns med datarengöring?
    a. Hur hanterar ni dessa utmaningar?
14. Hur gör ni för att säkerställa att ni bibehåller en korrekt representation av er data efter datarengöring, dvs att innebörden av datans betydelse inte ändras?
15. Hur ställer ni er till kostnaderna som uppkommer med datarengöring?

Ledningens stöd och engagemang

Litteraturen nämner också att en utmaning inom datakvalitet är att få ledningens stöd och engagemang.

16. Känner du att ni får ett bra stöd från ledningen i ert arbete med att upprätthålla god datakvalitet inom Big Data Analytics?
    a. Hur arbetar ledningen för att ge stöd till ert arbete?
    b. Är datakvalitet något som prioriteras och som det investeras i?

Övrigt

17. Finns det några fler utmaningar med datakvalitet inom Big Data Analytics som vi inte redan nämnt?
    a. Hur hanterar ni dessa utmaningar?

Bilaga 2 - Transkriberingsprotokoll T1

Verksamhet: F1 - Dagligvaruhandel
Medverkande: Respondent 1 (R1), Frida Carlsten (FC), Sofia Nyberg (SN), Emelie Uddenäs (EU)
Datum: 2020-04-29 **Intervjulängd**: 55 minuter

Rad	Person	Information (Svar/Fråga)
1	FC	Okej, så vi tänkte, vi ville först och främst bara höra med dig då i vilken roll du arbetar som idag?
2	R1	Ja, yes. Jag jobbar som... formella titel är Head of Data Science and AI. Och det innebär då ah ganska ordagrant då som titeln antyder att jag ansvarar då för på F1 moderbolaget eh så ansvarar jag för ett team som då är specialiserade för att använda just stora datavolymer, Big Data. Och oftast kombinerat med Machine Learning-teknologi och eh ja tekniska plattformer som lämpar sig för den typen av analys. Och så gör vi, försöker vi hitta mönster i data som tidigare inte varit möjligt då, med traditionella analysmetoder. Och allra helst så vill vi ju också automatisera smarta besult baserat på de mönster vi upptäcker då. Och vi stöttar alla dotterbolagen i sina projekt inom de områdena då. Så då är det F1 inom dagligvaruhandeln, ---, ---, sedan har vi också en dagligvaruhandeln kedja i de baltiska länderna också som heter --- och de jobbar vi inte mitt i än så länge, men något vi nog kommer göra framåt. Så det är de olika, ja de är ju olika branscher till och med vi jobbar med, men med den röda tråden då att vi använder Big Data och Machine Learning då och Data Science.
3	FC	Ja, men jätteintressant. Åh hur länge har du arbetat i den rollen du har idag?
4	R1	Ja, sen september 2018. Då startade vi hela funktionen, och jag fick möjlighet att bygga upp den från scratch då.
5	FC	Jaha, vad spännande. Okej… sen var det väl egentligen de här med arbetsuppgifter och ansvarsområden men de har du redan beskrivit så utförligt för oss eh... så de har vi koll på då. Då tänkte vi gå in lite på att det finns olika aspekter utav datakvalitet. Då har väl vi hittat att det finns lite olika i litteraturen, men att vi har valt att undersöka fem olika aspekter här utav datakvalitet som vi vill titta närmare på. Så jag kommer försöka förklara de olika aspekterna lite och sen undrar vi hur ni hanterar utmaningar inom dessa och om du ser några specifika utmaningar inom dessa områden.
6	R1	Mm okej.
7	FC	Eh ja, så du får hojta om de är något som är oklart där. Men då undrar vi till en början då vilka utmaningar som du ser kan finnas då inom konceptet då Information Completeness. Åh de vill säga då i vilken utsträckning som data saknas eller att den inte har tillräcklig bredd eller djup för den aktuella uppgiften, detaljnivån då att den kanske inte är tillräcklig eller avsaknad av data.
8	R1	Aa, vill ni ha på någon femgradig skala eller så eller vill ni bara ha ett resonemang kring det?
9	FC	Ah… vi tänkte väl mer ett resonemang kanske.
10	R1	Ah, men bra tack de blir lite lättare för mig också, men men men vill bara kolla så det inte blir för flummigt för er. Men man kan väl börja med att säga eh att vi absolut känner igen problemet och vi stöter på det ibland, men men jag skulle säga att det inte är ett av våra största datakvalitetsproblem för vi är lyckligt lottade då att ha väldigt mycket data på F1 och vi har ju också väldigt detaljerad data att vi har mycket uppgifter om våra kunder i och med att en väldigt stor andel av våra kunder väljer att identifiera sig när de handlar med vårt kundkort. Så eh då har vi mycket detaljer där och sen har vi ju också eh om man tar dagligvaruhandeln då särskilt så är det ju eh i sig väldigt detaljerat på

		transaktionsnivå. Att man handlar, våra kunder handlar i snitt i våra butiker någonstans mellan två till tre gånger i veckan. Och ah de är ganska många artiklar man plockar på sig liksom och då gör man ju massor av val liksom där det kan finnas de här spännande och komplexa mönster som kan ge oss hintar om vem den här kunden är och vad hon är intresserad utav förmodligen härnäst och så vidare. Så i detaljer ligger vi ganska bra till. Men med det sagt stöter vi på det ibland också, och då är det ofta, det är inte så att vi saknar data egentligen men vi har inte kommit igång med att spara undan det på ett --- sätt. Det händer, det händer ganska ofta.
11	FC	Okej, så då kan det vara en utmaning då alltså? Att man inte har lagrat datan menar du?
12	R1	Eh exakt inte på ett tillräckligt strukturerat sätt kanske. Man kanske har sparat det i en textlogg eller så. Det går ju ofta att ta ut också, men är oftast krångligare och tar längre tid. Och en annan klassisk dilemma, om vi har sparat det snyggt och prydligt i en databas men rensar historien varje månad så… att vi inte har, vi vill ju ofta ha ett eller två eller tre års historik när vi gör våra analyser och då händer det att vi saknar den ibland. Och då är det ju lite surt för då, okej, då ska vi slå på den här loggningen nu och vänta i tre år på att göra det vi vill, det är ju jättetråkigt då.
13	FC	Okej. Och hur, om ni stöter på ett sånt problem, som där exempel att man har sparat undan det i en textfil och att det tar längre tid, hur hanterar ni den typen av utmaning? Är det bara att man sätter sig och gör ett mer genomgående arbete för att få fram datan eller om man till exempel då har rensat historiken, finns det någonting, det finns ingenting att göra där eller?
14	R1	Aa juste. I första caset där om det är en textfil de är vi Data Scientist i regel ganska “hackiga” av sig, ganska duktiga på att lösa sånt. Men saknas det historisk så existerar det ju liksom inte, då är det kört. Då är det värre att det saknas historik än att det ligger ostrukturerat.
15	FC	Okej amen tack. Och då kan vi väl gå vidare där till nästa aspekt som är Data Accuracy. Då menar vi då att attributvärden är exakta rent språkmässigt, att datan är korrekt och pålitligt. Upplever ni att ni har några utmaningar inom den fronten?
16	R1	Aa men de har hänt, de är oftast väldigt svårt för oss i mitt team att märka. Att ah, vad ska vi ta, transaktionerna den här månaden går inte att lita på för då hade vi ett systemfel som skrev fel ner i databasen exempel. De är nästan helt omöjligt för oss att ta reda på, om inte det är uppenbart att det är fel exempel om vi förväntar oss en siffra och det står en text eller så. Då kan vi lista ut det, men står det en sjua istället för en nia så är det ofta jättesvårt för oss, i mitt team då. Men däremot de som jobbar nära, närmare data i sin verksamhet har ju oftast koll på sånt där. Och det händer, det var inte allt för länge sen vi fick ah, blev meddelade de från ett av våra bolag att vi har problem med ah data under de här månaderna så försök exkludera det från när ni gör analyser. Så det förekommer, men det är inte jättevanligt ska jag säga. Vi har en ganska bra nivå på, på, på Accuracy eller korrekthet i data tycker jag överlag. Men det är inte perfekt vi har våra brister ibland.
17	FC	Okej, om ni stöter på ett sådant problem. Hur hanterar ni det då?
18	R1	Ja, jag tänker spontant på två metoder som vi använder, och det beror i och för sig lite på case by case vilket som passar bäst. Men ett alternativ är helt enkelt att exkludera den delen av data i analysen. Och då också såklart tydligt beskriva det när man presenterar sin analys liksom att den här datan under den här perioden är inte liksom, ingår inte i modellen exempel på grund av det här och det här. Men det beror också lite på hur fel det är, men ett annat alternativ är att man inkluderar det, men att man då är extremt noga med att dokumentera. Och också när man presenterar data att man då pressar in en blasklapp i att den som tolkar den här analysen är medveten om att det finns felaktigheter.
19	FC	Man identifierar en felmarginal där då?
20	R1	Ah juste. Aa men precis. Det kan vara bra att man till och med hjälper den som ska tolka data i att försöka beräkna någon slags felmarginal där.
21	FC	Yes. Sen har vi då Data Currency och det handlar då om data är aktuell och up to date. Har ni några utmaningar med det?
22	R1	Aa.. ehmmm...jag skulle inte säga att vi har det inom så nämnbar grad i nuläget eh vi vi vi har tillräckligt aktuell data i princip i allt vi gör. Men det finns några så här framtida case som vi skulle kunna få utmaningar inom det här. Om vi skulle vilja göra mer analyser och prediktianor i realtid, att man till exempel i butiken när man går sitt kundvarv att man ska bli motiverad vid en viss plats med en ah något i stil med utifrån dina tre senaste artiklar som du har plockat på dig verkar det som att du handlar till köttfärssås och spaghetti och då har du tomatssåsen nu på din höger sida. Då måste vi data i

		realtid in liksom till vår plattform, och det har vi inte. Det är liksom dygnsladdningar oftast eh. Så då skulle vi kunna få det. Och vi kommer ju hamna i att vi vill göra sånt. Om inte exakt det exemplet jag tog, som var rent hypotetiskt faktiskt då, så kommer vi att komma till det förr eller senare. Och då är vi inte riktigt riggade för det utan då måste vi göra ett infrastruktursarbete också. Ta in data på ett nytt sätt då.
23	FC	Intressant. Eh använder ni er utav tidsstämplar på datan?
24	R1	Mm aa det gör vi.
25	FC	Ser ni något problem över att få tidsstämplarna överrensstämmande? Jag tänker beroende på vart man hämtar datan ifrån.
26	R1	Vi har inte stått på det som något problem än i och med att vi sällan använder tidsstämplarna till högre precision än timme.
27	FC	Nä okej. Och då har vi också Data Deduplication, vilket handlar om att det inte ska finnas dubbletter. Ser du att det finns några utmaningar inom det?
28	R1	Mmmm... det är ingenting som vi talar om som något stort problem i alla fall. Vi har i regel ganska bra skick på vårt data. Det mesta av det vi använder har passerat constraint kontroller i databaser och datavaruhus, just för att fånga upp dubbletter. Det kan vara när vi kopplar ihop data från två olika bolag exempelvis skulle vi kunna få… då gäller det att förstå till exempel att det kan vara en och samma kund som har ett id i ett av bolagen och ett annat id i det andra bolaget. Men vi gör väldigt lite sån analys där vi merchar data från flera bolag och det är främst och därför har vi inte stött på det så mycket. Men det finns där lite latent som en utmaning i alla fall. Sen tror jag det är hyfsat enkelt att hantera. Men vi har det liksom det potentiella problemet i all fall.
29	FC	Okej, och då när ni hanterar det, är det då genom constraints som du sa som ser till att det inte kommer dubbletter i databasen?
30	R1	Ah det är det traditionella sättet att hantera det och det är fortfarande det som präglar den mesta datan vi har. Vi kan koppla det till tidigare frågan där, angående aktuell data eller hur snabbt vi kan ta in data för att till exemplifiera, realtidsanalyser. Det hänger ju lite ihop med om du ska göra realtidsanalys, så har du inte tid att köra de här batchmässiga constraint eh delarna liksom. Utan då är det mera att skyffla in datan i en datalake utan att köra constraint kontroll. Och då måste man, då ställs det mer krav på analytikern än på data scientisten då att i sin kod att ta höjd för att det kan förekomma dubbletter och hantera det i koden dåra på något sätt. Och det kan man göra. Det viktiga är inte att glömma bort det. Ehh aah och få konstiga resultat.
31	FC	Yes. Okej. Åh Data Consistency är då den sista aspekten av datakvalitet som vi valt att undersöka. Då handlar det i vilken utsträckning som data presenteras i samma format och kompatibelt med tidigare insamlad data. Ser ni att ni har utmaningar i det?
32	R1	Mmm eh nej inte i någon större omfattning i alla fall. Det är väl ett tecken när jag behöver tänka efter… jag funderar på om vi ens har något sånt praktiskt exempel...Då är det inte vårt största problem i alla fall. Nä, nä, vi har bra ordning och reda av gammal hävd har vi jobbat länge med traditionell analys som BI-mässiga analys med att skapa rapporter och sånt där. Och då, det är ju lite signumet med den eran av analys att man just har mycket constraints, och lägger mycket krut vid inladdning av data. Att det är ordning och reda, single point of truth och sånt där. Sen den grenen av analys som mitt team jobbar med och representerar går lite emot det, och vill kanske hellre ha tillgång till rått data för att få det snabbare och alla detaljer i datat. Eh än att få det 100% kvalitetssäkrat.
33	FC	Ah, okej så ni hade kanske hellre haft som du sa lite råare för att datan inte skulle ha blivit för manipulerad?
34	R1	Aa men lite så. För det är ju ganska mycket intressant data som faktiskt filtreras bort i inläsning i datavaruhus typiskt för att man vid tillfället inte trodde att det var så intressant, alla attributen där. Så det är synd. Vi vill ju ha alla detaljer heh, i vårt jobb. Och tillbaka till det här med realtidsnära som möjligt, ehm vi tar det helst per ner på sekund eller vad det är. Om vi inte har användning för den detaljnivån så aggregerar vi hellre upp än att vi får det grovkornigt från början.
35	FC	Mm, så ni avgör hellre själva vad som är relevant för er att använda sen?
36	R1	Yes, yes.

37	FC	Okej, då tackar vi för det. Då har vi lite om ETL som är ett vanligt begrepp inom beslutsstödsvärlden som står för Extract Transform Load. Använder ni er utav ETL för att säkerhetsställa datakvalitet?
38	R1	Ja, det gör vi. Vi har... strukturen är då att vi har, våra datakällor som mitt team arbetar med finns i sitt ursprung i respektive bolag. Ex ---, ---, ---. Alla de bolagen har sina egna datavaruhus som är datagubbar från deras operativa system, transaktionssystem och lager och sånt där. Så där är det ju inte jag och mitt team som sköter ETL:en från de operativsystem utan det är ju respektive dotterbolag dåra. Men sen gör vi, en i mitt team tillsammans med ett systerteam som vi har på F1. Ska ni intervjua någon mer på F1 förutom mig?
39	FC	Nej tyvärr.
40	R1	Nej okej. Men jag kan tänka mig att det är ganska brett i alla fall. Men då är det ett annat team som jobbar med våran grupp, gemensamma datalakeplattform eller Big Data plattform som ehm som sköter ETL:en in till den då. Så då läser de från de här dotterbolagen datorvaruhus in till den här grupp gemensamma plattformen. Och ger data tillgänglig för bland annat mig team då att jobba med det. Så det är i… hur många steg blir det då, amen säg två steg dåra. Först i varje dotterbolag in i ett datavaruhus där, och sen från de bolagens datavaruhus in till den här grupp centrala gemensamma plattformen. Och i båda stegen sker det någon form utav ETL. Mer strikt constraint hantering i dotterbolagens inläsning och ganska lite constraint, om ens någon i… ah i det sista steget då.
41	FC	Okej. Skulle du säga att det finns några utmaningar med att använda sig utav ETL?
42	R1	Hmmm. Ah det beror lite på hur man definierar ETL dåra, men om man definierar det så som man har läst in data för analys traditionellt liksom, i mening. Då då är det lite tillbaka till det som jag sa innan liksom att vi ah att det finns risk att intressant data, liksom attribut på en entitet, en kund till exempel, filtreras bort. Det finns tillgängligt mer detaljer i det operativa systemet än vad som man tar ned när man läser in det i datavaruhuset. Så man tappar information, detaljer om information längs vägen. Det är väl den ena. Och den andra är det här… återigen att man fastnar i det här batch-tänket då, att man får uppdaterat data i typiskt nattkörning då. Istället för att få det flera gånger om dagen, eller till och med som en kontinuerlig ström av nytt data.
43	FC	Ja okej. Perfekt. Och då tänkte vi fråga lite mer om data samples. I litteraturen står det att även fast företag har stor kvantitet av data innebär det inte att samplet de använder sig av är tillräckligt representativt för hela populationen, alltså större är inte alltid bättre. Vi undrar hur ni gör när ni använder er av data samples vid analys, om ni använder er av det och om det finns några utmaningar i att välja ut samples.
44	R1	Ehm yes. Vi tar ofta ut ett mindre dataset tidigt när vi experimenterar med modellering typiskt då. Så då blir det otympligt att hantera de här jättevolymerna liksom. Utan då gör vi då som du beskriver med samples då. Eh och det finns ju, jag har ju duktiga statistiker i mitt team så dem kan säkerställa att vi kan ta ut en liksom ett 10%-sample från den stora, som är representativt med helheten och att vi kan lita på det. De resultaten vi når med de här samplet kan vi förvänta oss att vi kommer nå när vi skalar upp då.
45	FC	Du tycker generellt sett att ni får representativa sample helt enkelt?
46	R1	Ahh… det tycker jag… men det är väldigt viktigt att man har rätt kompetens när man gör det där momentet för det kan lätt bli lite fel och vilseledande om man inte riktigt vet vad man gör. Då kan man luras att tro någonting som inte är sant då kanske. Man får ett skevt sample helt enkelt. Det kan ju varar att man tar ut ett sample utifrån en... att man tar en tidsperiod till exempel i ett stort sample. Så kan ju det har varit en speciell tid som inte är representativt för helheten och sådär och då kan man ju gå bort sig. Man kanske plockar ut julveckorna eller så och så tänker man att alla 52 veckor under året säljer vi såhär mycket och dessutom verkar julskinka var något bra. Men det där… det är viktigt att man vet vad man gör helt enkelt.
47	FC	Definitivt. Och jag förstår själv att du kanske inte sitter och plockar ut samples, men du vet inte hur man går tillväga för att säkerhetsställa att man får ett representativ sample?
48	R1	Inte i detalj kan jag säga, men det finns statistiska modeller för det där och som också finns i amen typ Python-bibliotek som man kan använda sig utav och som de konfigurerar sig utifrån storlek och så vidare.
49	FC	Ja. Amen toppen tack så mycket. Då ska du få lite frågor från SN här istället.

50	R1	Ah okej okej ah.
51	SN	Då tar jag över. Ja, då går vi vidare lite till det här med felmarginaler. Och i litteraturen då poängteras vikten av att identifiera felmarginaler i ens data, då de menar att datakvaliteten aldrig kommer kunna vara perfekt. Hur ställer du dig till att det kan förekomma felmarginaler i den data ni använder er av?
52	R1	Aa. Aa men det är helt sant. Att… jag börjar svara i någon ände sen får du ställa följdfrågor. Vi brukar vara väldigt noga med att poängtera att det finns felmarginal i det mesta vi gör och presenterar. Det finns flera olika källor till fel. Dels är datakvalitet såklart, och särskilt om man har, extra lurigt brukar det vara när det är data som har någon form av manuell inmatning. Då kan man vara säker på att det är ganska hög grad av fel. Typ såhär eh vi har jobbat med något case där vi har kundtjänstdata där kundtjänsthandläggaren, medan de hade kunden i telefonen, klassificera vad för typ av problem och vad lösningen var och så där. Det var ganska taskig kvalitetet på det. Och det brukar vara det när det är manuellt då. Eh men det kan också vara andra fel som ah. Men vi försöker ofta titta in i framtiden och precisera vad som kommer att hända, vad som kommer sannolikt hända om vi utgår från dem mönster vi har sett i det historiska datat. Spolar fram bandet, och kör nuläget i modellen ah. Och därigenom, vad kommer troligtvis hända härnäst och sådär. Och det är ju liksom inte mer än en gissning, men det är en datadriven gissning i alla fall. Men det finns stora möjligheter till fel. Så det var ett väldigt långt och svävande svar kring att vi, ah vi brukar trycka på att det finns felkällor på olika sätt.
53	SN	Hur hanterar ni då dessa felmarginaler? Du var lite inne på att man poängterar mycket… eller hur hanterar man dem liksom
54	FC	Eller om man då gör uträkningar som du nämnde där innan till exempel eller ja.
55	R1	Aa men asså om man kan rätta till det så försöker vi att göra det, det finns ju. Och annars är vi noggranna med att presentera asså att beskriva för den som ska nyttja analysen att det finns, att man inte kan ta det 100% sant. Ehm men man kan ju rätta till saker, ehm ett vanligt datakvalitetsproblem är väl att det saknas… att det finns tomma fält och så där på vissa rader. Och då finns det ju tekniker hur man liksom rent statistiskt fyller i dem och så då ah och då är det samma sak att det är viktigt att man vet vad man gör då. Då efterkonstruerar man ju data så det är ju ganska, det kan vara ganska farligt, men det kan bli bra också liksom. SÅ det är en metod vi använder ibland i alla fall.
56	SN	Mm ja. Då tänkte jag också fråga, mäter ni erat datakvalitet på något sätt, och i såna fall hur?
57	R1	Nä det tycker jag inte att man kan säga att vi mäter datakvalitet så, på strukturerat sätt. Och det är nog också svårt liksom. Om man tänker sig data i numerär form till exempel, försäljning eller så det ska ju till Outlier-tal för att man uppenbart ska se att det är fel. Att något ligger väldigt avvikande från medel eller median eller så. Då kan man misstänka att det är fel, men det är inte säkert att det är fel faktiskt, det skulle kunna vara någon extraordinär händelse eller sådär. Ehm nä så vi går nog ganska mycket på att dem som vi jobbar tillsammans med på respektive bolag och verksamhet dem kan sin operativa världs så pass bra att dem vet i regel att här har vi, ah det här datat ska vi inte lita på 100% på för vi har haft IT-störningar eller vi har haft fel eller så. Så det är mest så vi får reda på faktiskt.
58	SN	Super. Då gör vi över till datarengöring och korrekt representation av data. Och där enligt litteraturen kan datarengöring användas för att lösa problem med saknade värden, syntaktiska fel i datainnehållet, syntaktiska skillnader mellan jämförbara dataobjekt, ursprunglig låg kvalitet av data, försämrad datakvalitet under lagring och avsaknad av metadata. Hur använder ni er utav datarengöring och vilka utmaningar anser du finns med det?
59	R1	Ehhh...Ja… Amen att en inbyggd utmaning med datarengöring är väl att man behöver förstå vad som är smutsigt dåra. Om man tänker rengöring, vad är dåligt vad är smutsigt data liksom, att kunna hitta det. Det är svårt, och särskilt om man är som vi, en central funktion som inte, vi jobbar inte operativt i de här verksamheters delarna som vi går in och gör enstaka projekt med. Så den här business förståelsen är en viktig del. Och eftersom vi inte har den själva måste vi få den genom teammedlemmar från den verksamhet som vi ska hjälpa. Så det blir väldigt mycket att vi ställer frågor och vi beskriver våra hypoteser och sen får dem rätta till det. Ofta blir det ju så aej men så är det ju inte, såhär kan du ju inte tolka det utan det är ju på det här viset liksom. Och ehm. Vad finns det mer… Du hade ju så många olika typer av rengöring eh. Jag är inte säker på att jag kan komma igång alla. Men det handlar ju ofta om sådär, jag tror du hade syntax som exempel ah vi får in data, vi får in data ett tal i en text-sträng till exempel sånt. Då behöver vi ofta omvandla det till en integer eller sådär. Det är väldigt vanligt datarengöringsmoment vi gör. Ah men de här med att fylla i blanka fält på ett lämpligt sätt. Att… plocka bort såna här outliers. Är ni bekanta med begreppet outliers?
60	SN	Nej.

61	R1	Det är då enskilda rader i ett dataset som avviker orimligt mycket från resten av raderna. Det kan vara ett dataaset från någon form utav, hur mycket varje kund har spenderat på -- den senaste månaden och sen så stöter man på en rad som står 300.000 eller så, då är det en outliner, den ligger utanför den förväntade skalan. Såna typer av analyser gör vi. Det är typiskt det första vi gör när vi bekantar oss med ny data, att liksom titta på, vilken range verkar varje kolumn ligga inom och sådär. Och då blottar det ofta i någon graf och då ser man exempel att ah den här raden 672 den sticker ut något enormt mycket. Det kan vara fel, oftast är det fel. Och då tar vi oftast bort den raden om vi har mycket data, så kan vi ta bort den för då, då blir den vilseledande när vi tränar Machine Learning på den.
62	SN	Ehm och hur gör ni för att säkerhetsställa att ni bibehåller en korrekt representation av er data efter datarengöringen? Alltså det vill säga då att innebörden av datans betydelse inte ändras.
63	R1	Ehmm ja hur gör man det… det är väl att… ja men det är väl... med det här oulier-exemplet att vi inte lägger vår förväntade storleksintervall på de här siffrorna alltför snävt så att vi plockar bort allt för många rader. Då kanske vi plockar bort ett helt segment av kunder som vi trodde var fel, men som faktiskt var korrekt. Och då blir resultatet därefter också, då har vi bortsett då från en riktig kundgrupp då. Det kan väl vara ett exempel då. Just de här med inputation, imputering kanske de heter på svenska, de här med att man fyller i siffror i tomma fält och sådär i efterhand, där kan man ju också gå bort sig och förvanska faktiskt. Om man inte gör det på ett rätt sätt.
64	SN	Hur ställer ni er till kostnaderna som uppkommer med datarengöring?
65	R1	Ehh… ja… kostnaden är väl i form av tid, oftast i vår värld då. Men vi tycker ju att det är viktigt, det är väldigt stora risker att inte kontrollera och rengöra datan där det behövs. Då riskerar man att ta fel beslut sen och det kan ju kosta mycket mer. Så tänker vi. Där kanske någon tycker det verkar onödigt, men vi brukar stå på oss i alla fall.
66	SN	Ja, och sista kategorin är då ledningens stöd och engagemang. Och i litteraturen nämns också att en utmaning inom datakvalitet är att få ledningens stöd och engagemang. Känner du att ni får ett bra stöd av ledningen för att upprätthålla en god datakvalitet inom Big Data Analytics?
67	R1	Ja men det tycker jag. Just datakvalitet tror jag inte att många i ledningen tänker jättemycket, men däremot så har vi väldigt tydligt stöd i att använda data för att ta beslut då. Och använda Big Data och sådär. Det är vår ledning väldigt bestämda i och ger väldigt bra uppvaktning för, i form av budgetar och att vi får anställa och så där. Sen är ju det snarare något kanske som vi eh… datakvalitet är möjliggörare som vi som jobbar med det känner till och som vi på något sätt bakar in i när vi beskriver vilken budget vi behöver ha till nästa år för att kunna fortsätta driva ett projekt baserat på Big Data. Så bakar vi in tiden som vi bedömer för de olika momenten och där datakvalitet är en del då. Och hittills så får vi okej på det vi föreslår och det är ett satsningsområdet på strategisk nivå för F1. Eh Big Data och Data Science och AI då. Så vi har tur på det sättet att det är ett prioriterar område.
68	SN	Tycker du att datakvalitet är något som prioriteras och investeras mycket i?
69	R1	Ehh… aa lite beroende på hur man ser det. Det är en naturlig del i vårt analysarbete egentligen. Sen sådär i varje projekt så har vi sådär oftast olika moment och de finns inbyggda i vår projektmetodikt i vår checklista för de olika stegen vi går igenom så ska vi alltid kontrollera och säkerhetsställa kvaliteten på datan. Ett annat sätt att attackera det där är att göra något kanske större och mer renodlat projekt för att förbättra datakvaliteten mer mot källorna, i datavaruhusen och sådär. Något sånt projekt, vi har inte gjort något jättestort sånt projekt ehm… där man tar mer ett samlat begrepp på datakvalitet kanske på flera ställen samtidigt. Men vi har ännu inte riktigt sett det behovet än, vi tycker att det fungerar, att vi löser det för varje projekt som vi gör.
70	SN	Ja. Som en sista fråga då tänkte vi fråga om du tycker det finns några flera utmaningar med datakvalitet inom Big Data Analytics som vi inte redan har nämnt?
71	R1	Oj ja jag tycker ni har täckt det bra. Men ska funderar några sekunder… Nä jag kommer inte på några nu men om jag gör det så kan jag ju maila dem dig, jag har ju din mailadress. Jag tycker det bra frågor. Svåra men bra.
72	SN	Vi förstår det. Men det var allt från oss. Tack så mycket för du ställde upp.
73	R1	Tack själva och lycka till nu!

Bilaga 3 - Transkriberingsprotokoll T2

Verksamhet: F2 - Konsultbyrå. Respondent arbetar på ett industriföretag.
Medverkande: Respondent 2 (R2), Frida Carlsten (FC), Sofia Nyberg (SN), Emelie Uddenäs (EU)
Datum: 2020-04-29 **Intervjulängd**: 40 minuter

Rad	Person	Information (Svar/Fråga)
1	FC	Vi vill bara börja med att få höra lite om dig och veta i vilken roll du arbetar som idag?
2	R2	Yes, jag är anställd som Junior Data Scientist och det är väl antagligen det jag fortfarande är det är ett år sedan jag började. Jag har egentligen suttit med ett och samma projekt hela tiden, jag började med ett nytt nu precis. Men då är det jag tillsammans med --- som har haft ett projekt från början till slut och egentligen gått från allting som Data Processering, modellering och Deployment av den här modellen.
3	FC	Okej och hur länge sa du att du har arbetat med detta, med Data Science?
4	R2	Det är ett år sedan jag började på F2 och då fick jag ganska precis uppdraget på Företaget också, så ett år skulle jag säga.
5	FC	Okej. Nu är det som så att vi har läst i litteraturen att det finns lite olika sätt på att definiera datakvalitet. Det finns lite olika aspekter rättare sagt som kan användas för att definiera datakvalitet. Så jag kommer att ge en liten kort förklaring till de sex vi har plockat ut och valt att tittat närmare på. Så kommer jag fråga dig om du kan förklara vad du ser för utmaning inom det området och hur ni i sådana fall arbetar för att hantera det när ni stöter på den typen av utmaningar.
6	R2	Yes!
7	FC	Så den första aspekten heter Information Completeness och det handlar om i vilken utsträckning data saknas, eller att den inte har tillräcklig bredd eller djup för en aktuell uppgift. Så vilka utmaningar tycker du att det finns just inom Information Completeness och hur skulle ni hanterat dem?
8	R2	En sak direkt jag tänkte är att ofta ska man göra, ja det vi har jobbat med nu är att vi ska försöka prediktera hur stor risken är att ett projekt ska överskrida budgeten. Det är så mycket som kan gå fel i det, det är väldigt väldigt svårt att få en bra modell på det. Datan vi utgår ifrån är ett sådant rakt projekt och det är utföringsdag, ehh ja hur stor budget det har varit till projektet, hur budgeten är upplag och sådär då. Det är ganska lite information som ska kunna avgöra en så komplex sak, om det skulle kunna vara ett exempel kanske. En sak vi har gjort därför för att försöka tackla det problemet är då att vi har haft Ensemble modeller. Jag vet inte om det blir för tekniskt kanske, men att man har massa olika modeller som… men så att man har massa olika gissningar på prediktionerna så tar man ett medelvärde generellt.
9	EU	Vad sa du för modeller? Förlåt att jag avbryter.
10	R2	Ensemble modeller
11	EU	Okej, kan du bara förklara lite?
12	R2	Ahh det är rätt enkelt. Jag vet inte hur mycket ni kan om modellering och sådär med modeller, men om man tänker sig att man har någon slags logistisk regression. Så istället för att ha en modell som bara gissar på en prediktion vad det kan bli, det det också är vag data så är det ganska lite att gå på. Så har vi haft flera hundra modeller istället, så vad du än gör i prediktionen så blir det ganska olika

		allihopa. Men sen tar man medelvärdet av alla utfallen och har det som gissning istället. Ungefär som att man låter hundra personer gissa på en sak, så tar man medelvärdet för att få en lite mer robust gissning.
13	FC	Ja okej! Då går vi vidare till nästa aspekt. Det är Data Accuracy och det handlar då om att datan är korrekt och pålitlig, certifierad att den är exakt rent språkmässigt och har en korrekt detaljnivå.Vi undrar då vilka utmaningar det finns inom Data Accuracy, som du stöter på i ditt arbete och även hur du skulle hantera dem?
14	R2	Ja, ofta är det ju att man har felaktiga värden i datan helt enkelt. Man ser någon data i någon kolumn och så ser man en budget som är negativ och det känns ju jätteskumt. Det vi brukar göra i första hand är att snacka med någon expert på datan, som kan datan utan och innan, och fråga varför är det så här och vad beror det på. Ibland får vi svaret att amen det kan så ut sådär det beror på detta blablabla eller så är det bara att någon har fyllt i något felaktigt. Jag menar ibland står det när ett projekt startades år 9999 och då tänker man direkt att det är någonting som man bara skrivit för att fylla i någonting. Men i första hand är det att prata med en expert som kan datan utan och innan, då kan dem till och med i vissa fall fylla i luckor och sånt här ifall de vet vad som ska stå där istället.
15	FC	Okej!
16	R2	Det skulle jag säga är nummer ett. Jag har skrivit upp lite små anteckningar här så jag ska se om jag kan hitta något där i. Ibland om det saknas data som vi hade nu till exempel för ett tag sedan i ett projekt, att man visste hur budgeten såg ut efter. De hade först en rad, det skulle funnits en rad för hur projektet såg ut i början vad det hade för budget och sånt till exempel men sen så fanns inte den raden utan det fanns bara själva uppdateringarna senare. Så vi visste inte utgångsläget utan vi visste bara vad som kom senare, men då så kunde dem se lite smart i kolumnerna att det fanns något som hette Previous Value till exempel. Då kunde man lätt återskapa den förra raden med hjälp av den senare raden. Så lite sånt kan man också hitta, småsaker i vissa kolumner på datan.
17	FC	Som kan hjälpa en liksom och göra den...
18	R2	Precis, ja så man får leta efter ledtrådar här och där och se vad man kan göra för att återskapa det som saknas och så.
19	FC	Ja okej, amen super! Då är nästa aspekt Data...
20	R2	Kan jag bara ta en sak till? Det finns också om man kan kolla ett värde som saknas, ett numeriskt värde ofta då, så kan man ofta köra lite interpoleringar och sådär också så man kanske tar något medelvärdesinterpolering någon tidsseriedata, det finns olika metoder för det.
21	FC	Okej, då var nästa aspekt sen Data Currency. Det kan även benämnas som Timelessness. Det handlar egentligen om hur aktuell datan är eller huruvida den är aktuell ens. Så då är frågan här vad du anser att det finns för utmaningar inom Data Currency, då med aktuell data, om det är någonting ni stöter på?
22	R2	Mmm, nu har jag bara jobbat ett projekt i princip så jag tar väl det igen. Vi kollar ju då på att börja träna en modell för att se om projekt i framtiden ska överskrida budgeten så har vi kollat på avslutade projek. Då har vi ju haft ganska mycket problem med att vi har få samples där. Så det har väl kanske ja varit ett par hundratals projekt att utgå ifrån kanske, men man kan så klart få fler projekt om man går längre bak i tiden. Vi gick årtal tio år tillbaka i tiden, för att kunna få så pass mycket att vi kan jobba med det. Men då har man ju det problemet att tio år bak i tiden såg saker och ting annorlunda ut och man hade andra förutsättningar när man gjorde projekt. Så det kan vara helt andra saker där som går fel än som går fel i nuläget så alla samples kan vara lite olika beroende på när man kollar på dem. Så även om vi får mer när vi går bak så kan det bli knasigt också.
23	FC	Ja och hur gjorde ni då för att hantera den utmaningen? När ni stötte på att ni bara kunde gå tio år tillbaka i tiden och att ni inte vill gå ännu längre. Var det helt enkelt att ni valde att avskärma det där för att ni mena att längre tillbaka var det inte aktuellt eller vad det okej att bara använda de tio åren då?
24	R2	Vi kände väl att vi tog och kapade av det där för att det kändes för långt bak annars. Det kändes inte så aktuellt längre då och vi nöjde oss med de samples vi fick.

25	FC	Jag förstår okej. Nästa aspekt heter Data Deduplication och det handlar om att det inte får finnas dubletter eller data som refererar till samma riktiga entitet så att säga. Så man ska därför ersätta alla dubletter så det bara finns ett korrekt värde. Då undrar vi ifall, vad du har, stött på för utmaningar inom det?
26	R2	Lite har man ju sett här och var, typ med sånt här att closed date kan vara två olika datum av någon anledning. Men då är det återigen där att man snackar med någon som kan datan väldigt bra, för att försöka förstå varför det finns två och vilken av dem som är den vettiga. Ofta med just datum kan man ju tänka sig ju att om man har ett closed date att man tar det som är senast då för att dem kanske har stängt ett projekt vid ett tillfälle och sedan insett att ojn det var inte färdigt, så stänger dem det igen senare. Då är det ofta att man tar det senaste i det fallet. När det gäller annat än datum så vet jag inte om jag har haft så mycket med det faktiskt.
27	FC	Nä, har ni haft på något sätt dubletter av annan form? Jag tänker om det skulle kunna bli dubletter i data som hämtas in från olika system? Exempelvis i form av att de håller data om samma entitet.
28	R2	Ehh de har vi nog faktiskt inte, jag tror faktiskt inte det.
29	FC	Nä okej tack! Men då har vi åtminstone ett exempel där och du förklarade hur ni hanterade det ja. Sista aspekten Data Consistency handlar om i vilken utsträckning data presenteras i samma format, som då tidigare data, att det är kompatibelt med tidigare hämtad data. Tycker du att ni har haft några utmaningar inom den biten?
30	R2	Där har det nog varit ganska bra faktiskt. Det har varit samma mall som har använts i alla dem här tio åren känns det som. Så att där har det nog inte varit några problem.
31	FC	Nä
32	R2	Däremot har vi haft lite problem när vi hämtat data om… Vi har det här datat vi jobbat med i detta projektet som är själva implementationen av projektet, men sen har man också en fas tidigare som är när man skulle binda det här projektet och sälja in att företaget skulle få göra det. Åhh med den datan kan man tänka sig att varje projekt där i början skulle bli ett riktigt projekt, att det skulle kopplas till ett projekt som faktiskt har implementerats sen. Så det var två olika faser men där har man inte alls lyckats koppla dem tillsammans, för då är det verkligen att så fort man är förbi den här försäljningsfasen så har dem bara släppt det och gått vidare här och så kan man inte riktigt koppla försäljningsfasen till nästa fas. Jag vet inte om det är aktuellt till frågan kanske?
33	FC	Jo men det kan väl liknas med det precis! För då har det ju inte riktigt varit att man har kunnat fortsätta med tidigare data eller?
34	R2	Nej precis, för det är inte samma data exakt men det är ändå samma projekt.
35	FC	Mmm ja, intressant! Så då har man inte riktigt… Har man haft något sätt att hantera det på eller hur har man valt att hantera det? Har man bara valt att gå vidare sa du?
36	R2	I det fallet så har vi faktiskt till och med haft någon person, någon stackare, som manuellt kopplade ihop dem. Det var kanske en sådär femtio stycken eller någonting så han fick sitta där och försöka lista ut vilken som tillhörde vilken och sådär. Men nu har dem bara släppt det tills vidare för att det var så svårt. Däremot har vi också sagt till nu att det har varit ett problem och dem kommer kolla på att försöka skapa den här kopplingen i framtiden.
37	FC	Okej, då har vi en fråga om ETL också. Jag vet inte om du känner till ETL, men det är ett begrepp som används inom beslutsstödsvärlden och det står då för Extract, Transform och Load.
38	R2	Mmm
39	FC	Det känner du till då antar jag?
40	R2	Lite grann, inte jättemycket.
41	FC	Nä okej, detta är då vad vi har läst oss till ska vi säga enligt litteraturen och då kan man använda sig av till exempel när man hämtar data från ett system och ska överföra det till ett annat eller från en databas in till ett Data Warehouse eller liknande. Så då är egentligen frågan om du vet ifall ni använder er utav ETL? Men det kan ju vara lite svårt om du inte sitter med den biten.

42	R2	Ja det kan jag nog inte riktigt svara på tror jag.
43	FC	Nä okej, det är helt okej vi förstår det.
44	R2	Det känns som att alla har sitt sätt att hämta data på så det är lite… lite ja där kan väl folk göra lite som de själv vill. Ibland så tar man en SQL Query och ibland får man en fil från någon, det är väldigt olika.
45	FC	Okej, jag förstår. Men då hoppar vi vidare till nästa som handlar om urval utav data samples. Enligt vad vi har fått fram av litteraturen så fastslås det att även om företag har stor kvantitet av data så innebär inte det att deras sample nödvändigtvis är representativt för hela populationen. Det vill säga att större är inte alltid bättre. Då är vi intresserade just utav hur man väljer ut de här samplesen för att säkerhetsställa att de ger en representativ bild utav hela populationen. Så vår första fråga till dig är om ni använder er av data samples när ni gör analyser?
46	R2	I vårt fall antar jag att det i så fall skulle innebära att varje sample är ett projekt vi jobbat med?
47	FC	Jaa, till exempel om man då har en massa data och ska göra en analys så kanske du väljer att ta ut ett dataset då, alltså en liten del bara, som du kanske kör en testanalys på.
48	R2	Ofta blir det ju… Jag har lite svårt att se vad som kopplas bara. Jag är inte helt med på frågan faktiskt.
49	FC	Nä, tror du att du kan förklara det bättre Emelie?
50	EU	Amen, när man använder sample typ... För om man skulle kört all den här datan som man har då så kan det ta ganska lång tid att bygga modeller eller analysera det så man använder väl, alltså jag kan tänka mig att man använder statistiska uträkningar för att välja ut ett mindre dataset, men som ändå representerar hela datasetet. Men det kan ju liksom uppkomma lite utmaningar, om ni använder er av det så kan det ju uppkomma problem med att göra det också. Till exempel då att det kanske inte ger en fullständig bild av hela populationen.
51	R2	Okej, ahh det har vi gjort lite grann att man har plockat ut… Till exempel i det fallet där vi har ganska få samples att jobba med och då måste man ha en viss del av samplesen för att träna modellen på och en viss del för att testa den på. Sen blir det ganska lite träning och lite testing eftersom man delar upp det och redan har lite från början. Det vi har gjort då är att vi har plockat ut… Ja vi har gjort flera körningar, så ibland tar vi ut denna delen till träning också tar vi en del till testing. Sen gör vi om samma sak igen fast vi tar en annan del till träning och en annan del till testning. Så man på något sätt använder all data till både träning och testning. Kallas för ---validering och då får man se till att man har ungefär lika stor proportion av olika data i varje sample så det blir lika för varje gång.
52	EU	Mmm, men det skiljer sig inte? Känner ni då att det blir liksom ungefär samma resultat oavsett vilket sample ni väljer ut eller? Ni känner aldrig att det är något problem och tänker oj det här blev inte alls…
53	R2	Jo i vårt fall skiftar det ganska mycket eftersom det är så små datasets. Så att ett projekt som är väldigt konstigt om det hamnar i ett träningsset eller testset så kan det påverka ganska mycket på utfallet på modellen. Det kan växla rätt mycket så fall. Men jag tänker har man väldigt stort dataset så borde det inte vara något problem utan då borde det bli ganska konsekvent.
54	EU	Men om det då skiljer sig så pass mycket, hur hanterar ni det då? Använder ni det ändå eller gör ni om det eller?
55	R2	Vi får helt enkelt använda det ändå för att ja det är väl lite därför man har den här ensemble modellen till också så man gör många gissningar och tar ett medelvärde.
56	EU	Mmm okej, så ni jämför de olika samplesen med varandra då och försöker hitta något mellanting?
57	R2	Ja alltså du måste alltid ta medelvärdet på själva prediktionen sen efter modellen.
58	FC	Men efter att ni då gjort ett urval där som ni testar och tränar som du sa, gör ni någonting för alltihopa sen? Eller gör ni det bara uppdelat i olika urval, i olika samples, eller gör ni det på all data sen också?
59	R2	Det gör vi sen också när vi väl ska använda modellen. Vi gör den här uppdelningen bara för att träna modellen och sen när vi väl ska använda modellen på riktigt så tar vi all data till träning bara och sen så brukar vi inte utvärdera den utan då har vi all data för att anpassa modellen.

60	FC	Yes, men då ska du få lyssna lite på Sofia här istället.
61	SN	Nästa del handlar då om felmarginaler. I litteraturen poängteras vikten av att identifiera felmarginaler i ens data, då de menar att datakvaliteten aldrig kommer kunna vara perfekt. Så hur ställer du dig till att det kan förekomma felmarginaler i den data som används?
62	R2	Jag brukar alltid tänka att… Eller en vanlig fråga jag brukar ställa är om en kolumn i ett dataset är ifylld manuellt av en person. Om det är manuellt ifyllt brukar det alltid vara massor av konstiga fel som inte går att förklara riktigt och det är svårt att reda ut vad felen beror på. Det kan bara vara någon som skrivit fel, ett typo, eller inte brytt sig om vad den skriver in bara. Men i dem automatiskt ifyllda kunduppgifterna är det i bästa fall alltid rätt, men det är ju sällan det blir så.
63	SN	Så det är oftast mer liksom rätt när det är automatiserat?
64	R2	Ja det brukar ofta vara mer stabilt då eller om man bara fyller i… Om någon annan ska fylla i en kolumn så kanske man har någon sorts check att det den fyller i faktiskt kan vara trovärdigt. Jag menar om man skriver in ett årtal som är 9999 så kan man ha en check som säger att det där är inte rimligt och så får man fylla i igen. Vad var frågan nu?
65	SN	Jag tror den var hur ställer du dig till att, eller liksom hur ställer du dig till att det kan förekomma felmarginaler i datan. Men som du säger det är väl att det händer liksom.
66	R2	Ja det känns som att det finns nog inget dataset utan fel, det är vad jag har sett iallafall utan det är alltid massa… Ofta blir det bara att man tar bort dem som man inte har någon förklaring till, att man bara trycker drop och så försvinner dem. Men det är ju inte alltid det bästa sättet att bara göra så.
67	SN	Nej exakt och det är fråga, andra frågan då. Hur ni hanterar den här utmaningen med felmarginaler.
68	R2	Det blir ju ofta att man bara tar bort dem tyvärr. Vissa går bara inte att göra något åt utan det är kört helt enkelt. I andra fall så får man då fråga experten igen och se vad dem kan säga om det.
69	SN	Aa jag förstår.
70	FC	Om ni då vet att ni har identifierat… Alltså även om det kan finnas vissa felmarginaler då i datan är det så att man kan acceptera det och ändå använda sig utav datan och att man är medveten om att där kanske är en viss felmarginal? Eller blir det alltid då att man tar bort den? Eller måste du försöka få datan perfekt, även om det då kanske inte går innan ni använder den? Så att man då tar bort om det är tveksamt eller kan man acceptera att det finns en viss felmarginal med?
71	R2	Amen ofta har vi nog en del fel också, det stämmer. Vad vi brukar göra är att vi gör ett histogram över felen och så kanske vi säger att dem som är inom 80% korrekt eller något sånt där, så man har någon slags gräns med tröskelvärden. Så allt som är inom det här kan man ta med i analysen och sen allt annat som är sämre skippar man bara. Det gör vi ganska ofta också.
72	FC	Okej, så då gör ni ett val där över vad ni tar bort och vad ni behåller. Då när ni är medvetna om att där är en felmarginal gör ni något vidare med det sen eller då kör ni bara analysen? Eller tar man det i åtanke vid analysen också?
73	R2	Det kan hända att man tar med det också i analysen. Man brukar alltid spara med det iallafall så man har kvar det, så man inte bara tar bort den informationen och tror att alla är perfekta. Utan man sparar alltid med det värdet efter. Så att i analysen när man väl kollar på resultatet kan man då kolla och ta bort de här som var dåliga då och bara kolla på de som man vet är trovärdiga. Sen om man vill ha lite mer så kan man ta med dem som var lite halvdanna också.
74	FC	Ja okej! Så att man är medveten om det när man kollar på analysen.
75	R2	Ja
76	SN	Ja och vet du om ni mäter er datakvalitet och i så fall hur gör ni det?
77	R2	Inte vad jag har hört om iallafall. Jag vet inte om det är något sånt där officiellt att man har något på kvaliten så.
78	SN	Nä, ville du säga något Frida?

79	FC	Jag skulle väl lägga till lite där att vi läst i litteraturen att det är väldigt svårt att mäta datakvalitet just för att det består av många olika aspekter och att det kan behöva sättas i förhållande till vad det ska användas till. Vi tyckte att det var intressant att veta i och med att det är en så pass stort utmaning att försöka mäta kvaliteten. Så om det fanns något sätt som ni gjorde för att avgöra hur bra kvaliteten är.
80	R2	Jag tror det bara är mer att man snackar internt inom vår Data Science grupp man snackar om att den här datan är bra kvalitet. Men det kan hända att det finns något bakom som jag inte vet om, det är mycket möjligt. Man brukar ju ha någon som är dataansvarig och dem kanske sitter med sådana typer av saker.
81	SN	Ja, då går vi vidare till datarengöring och korrekt representation av datan. Då enligt vad vi läst i litteraturen så kan datarengöring användas för att lösa problem med bland annat saknade värden, syntaktiska fel i datainnehållet, syntaktiska skillnader mellan jämförbara dataobjekt, ursprunglig låg kvalitet av data, försämrad datakvalitet under lagring och slutligen avsaknad av metadata. Så använder ni er… Hur använder ni er av datarengöring och vilka utmaningar anser du det finns med det?
82	R2	Jag brukar ofta göra så att man funderar lite på vad man vill ha i slutänden. Vad man vill ha för typ av data, hur man vill att det ska se ut och sådär och sedan utgå lite från det. Så att är det någon kolumn man inser att den här kommer inte jag behöva kan man bara släppa den direkt då. Så slipper man tänka på att rengöra den, så att man lite begränsar sin arbetsbörda. Sen dem som man väl har, om man inser att någon kolumn är väldigt viktigt får man lägga fokus på den och kanske snacka med någon expert och försöka göra rent och kanske lägga någon dag på att bara fixa den kolumnen. Få den att bli användbar.
83	EU	Men innan datan kommer till er, har den gått igenom någon datarengöringsprocess då? Eller är det ni som också gör det då?
84	R2	Det är nog lite olika. Det finns något team som sitter mycket med sånt också. Åhh då ät tanken att de ska ge oss den data som är ganska rengjord, men sen är det väl en lite vag gräns också med vad som går att göra rent i datan och att processera den till det man behöver. Så ofta blir det att vi sitter mycket med data cleaning också.
85	SN	Så finns det några… Är det några specifika utmaningar med det som du kan komma på i huvudet?
86	R2	Det är väl lite allt det vi snackat om med avsaknad av värden och felaktiga värden och allt det där. Det är mycket det som sätter käppar i hjulen när man vill komma fram. Sen ofta dyker det upp i en kolumn där det ska stå en siffra så kan det dyka upp ett bokstavsvärde eller något sånt där. Det kan vara väldigt konstiga saker som händer.
87	SN	Och ni hanterade det genom att… Ja okej hur gör ni för att säkra att ni bibehåller en korrekt definition av er data efter datarengöringen? Alltså det vill säga att innebörden av datans betydelse liksom inte ändras efter datarengöringen.
88	R2	Nu är jag inte helt säker, men du kanske menar att om man tar ut… Säg att man hämtar data från en databas och man rengör den och sådär och fixar. Du menar då att det kanske uppdateras då där man hämtar datan ifrån sen eller hur tänker du?
89	SN	Ja men det är väl lite det här kanske att om man skulle skriva in nya värden exempelvis om det saknades värden eller så. Det är väl det vi…
90	FC	Ja alltså det är väl ett exempel. Det finns ju många olika processer för datarengöring men till exempel det exemplet som Sofia tog där att man fyller på värden i fält där det saknas data, då kan det ju va att du på så sätt manipulerar datan till att den faktiskt inte har samma innebörd riktigt längre. Att det inte längre är någon rättvis representation av ursprungsdatan. Att det har fått liksom en annan innebörd.
91	R2	Okej. Jaa så ni menar att man ska hantera att den inte ska ändras utan bara behålla sig som den är eller? Jag är inte helt 100% där.
92	FC	Alltså mer att innan du då kan gå vidare och göra analysen så vill du ju inte ha fält då som saknas så då kanske ni fyller på med värden och då kan ju det råka… Om man inte gör det korrekt eller noggrant så kan det ju leda till att man fyller på med felaktiga värden eller sådant som gör att det sedan blir en skev analys. Så frågan är vad man gör för att säkerhetsställa att det inte sker. Då var det här med avsaknad av värden ett exempel, det finns ju många olika sätt eller processer vid datarengöring.

93	R2	Om man ska göra en analys på datan, om man ehh... så brukar vi göra så att vi tar liksom som ett snapshot av datan och --- och sen vill man gärna behålla det som det ser ut som --- låtit den uppdatera sig i regel utan under analysen så har man den datan man har och sen låter man det vara dem ändrar inte på den rådatan utan man sitter till och med skrivskydd på datan så att det inte ska kunna ändras på något sätt för att man ska kunna reproducera sitt resultat som man har fått i analysen. Åh även om man då skulle vilja få ett nytt värde så kanske man gör det på ett annat ställe i sådana fall, men låter den rådatan vara kvar som den är bara för att kunna få samma resultat igen.
94	FC	Ja okej
95	R2	Jag vet inte om det var svar på frågan kanske?
96	SN	Jo det var det.
97	R2	Ja okej
98	SN	Ja, hur ställer ni er till kostnaderna som uppkommer med datarengöringen?
99	R2	Det känns som att Företaget är lite så att de har pengar så att dem tänker nog inte så mycket på det vad jag vet. Utan att man får lägga så mycket man vill på det känns det som. Ibland vill dem väl pusha på lite grann för att de vill att det ska gå snabbt, men det är nog mer för att man vill ha resultat än att det kostar tid, kostar pengar på rengöringen. Jag skulle nog säga att kostnaden är inget problem det är mer att man vill ha resultaten fram.
100	SN	Ja och då sista kategorin handlar om ledningens stöd och engagemang. Och litteraturen nämner också att en utmaning inom datakvalitet är att få ledningens stöd och engagemang. Känner du att ni får ett bra stöd från ledningen i ert arbete med att upprätthålla god datakvalitet inom Big Data Analytics?
101	R2	Det känns som att vi är nog inte riktigt ansvariga för kvaliteten på datan på det sättet, utan det är mer att vi använder den data som finns. Ehh så attee...
102	R2	Var tanken att vi skulle varit ansvariga för datakvaliten?
103	EU	Nä men jag… alltså typ mer att dem väljer att... nä men att de ser processen med att få bra datakvalitet, så att man kan använda den datan till så mycket som möjligt, att de ser det som en viktig del i ert arbete och kanske ger er, alltså så att investerar i det både tid och pengar.
104	R2	Jag kan nog säga att det är lite för mycket fokus på att få ut resultatet så mycket som möjligt. Sen är det inte lika mycket fokus på att få ett bra resultat utan det är mer att vi får ut ett resultat, så man kör lite mer på det här med kvantitet över kvalitet. Det kan jag tycka är lite synd att man hoppar gärna över lite steg och sådär och kanske lägger lite mindre tid på kvaliteten på datan, istället för att försöka få fram en modell som funkar någorlunda bra istället för att få fram en som funkar väldigt bra.
105	EU	Så då kan man väl säga egentligen att dem stöttar mycket till det här med analyserna och det, att få fram resultat av modellerna. Att ni får mycket stöttning i den processen men mindre i alltså processen med att få fram datakvalitet eller?
106	R2	Ja men så skulle man kunna säga tycker jag.
107	EU	Mm ja!
108	SN	Ja då har vi ställt alla eller jo sista frågan då och det var egentligen bara att vi ville kolla om du själv kommer på några mer utmaningar med datakvalitet inom Big Data Analytics som vi inte redan har nämnt?
109	R2	Förlåt vad var det där sista ordet?
110	SN	Nä men vi tänkte bara kolla om du kommer på några själv bara sådär i huvudet, några fler utmaningar som med datakvalitet inom Big Data Analytics som vi inte redan har nämnt.
111	R2	Ehh, inte på rak arm tror jag inte. Men jag skulle ju kunna återkomma sen om jag kommer på något mer kanske.
112	SN	Ja men det skulle vara jättesnällt!

Bilaga 4 - Transkriberingsprotokoll T3

Verksamhet: F3 - Revision
Medverkande: Respondent 3 (R3), Respondent 4 (R4), Frida Carlsten (FC), Sofia Nyberg (SN) och Emelie Uddenäs (EU)
Datum: 2020-05-05 **Intervjulängd**: 44 minuter

Rad	Person	Information (Svar/Fråga)
1	SN	Okej,vad bra. Vi tänkte först höra med er i vilken roll ni arbetar som idag?
2	R3	Jag heter R3 och jag har arbetat på F3 i 2.5 år. Jag pluggade också i Lund faktiskt fast jag pluggade Civilingenjör inom Datateknik. Och vi arbetar i en ganska dynamisk vardag med lite olika projekt så man har gjort lite sådär något med en chatbot, något med robotik, något med Machine Learning så att vi har snuddat dem här beröringspunkterna på lite olika sätt tror jag. Ehm men primärt så arbetar jag inom den sfären av AI-instrument och Machine Learning och Robotik.
3	SN	Mm, Vad heter eran position just nu då liksom?
4	R3	Management Consulter men vi jobbar ju mer åt det tekniska hållet än en traditionell Management Consultant. Sen Management Consulting har ju varit fram tills, men ganska sent ändå, ren strategi-consulting. Men i och med att det digitaliseras mycket så behövs det eran kompetens då och våran kompetens för att faktiskt förstå sig på IT-instrumenten som ett bolag ska implementera. Så det behövs expertis inom det området. Och det brukar väl inte röra sig om längre utvecklingsprojekt, utan det är väl mer att man visar bolag vad dem och hur dem kan använda sig av teknikerna och visar liksom hur man drar mest nytta av dem och sådär. Så det brukar röra sig om sådär uppstartsprojekt. Kanske att man utvecklar i några månader och sen tar företaget över när dem kan behärska teknikerna, så att det är väl så vi arbetar.
5	SN	Yes, och R4?
6	R4	Jag har jobbat på F3 samma tid som R3, som du sa då R3 så arbetar vi sen Augusti förra året. Eh kommer tidigare från Linköpings Universitet där jag läste Civilingenjör inom Industriell Ekonomi med inriktning Datateknisk inriktning. Därmed hamnade jag inom detta. Så vi gör väl lite samma sak jag och R3. Vi varvar rena utvecklingsprojekt som sagt med att stötta företag med deras IT-relaterade strategi frågor, med datahantering osv.
7	SN	Aa men super. Vill ni tjejer ta över då?
8	FC	Ja, hör ni mig nu? Snyggt. Okej, men då kör vi igång med lite intervjufrågor. Och det vi har gjort nu när vi har kollat på datakvalitet är att det finns många olika definitioner av vad datakvalitet är för något. Och vi har valt att fokusera på vad blir det, fem stycken olika. Så jag kommer fråga lite mer kring dem aspekterna av datakvalitet specifikt och så får ni svara på om ni ser att ni har stött på några utmaningar inom dem här aspekterna och i sådant fall hur man har hanterat dem. Så den första utmaningen då är Information Completeness och det handlar om i vilken utsträckning data saknas eller inte har tillräcklig bredd eller djup för en aktuell uppgift. Vad ser ni att ni har stött på för utmaningar inom detta. Är det något ni har upplevt eller brukar råka ut för?
9	R3	Ja, alltså så här. När man utvecklar och när man använder större datamängder och baserat på den datan som ska ut på olika uppgifter och då sätter man vanligtvis upp en threshold på sina modeller eller sitt hjälpmedel i form av AI då. När den är så pass osäker då så att man kanske ska ringa i

		varningsklockor. Och det är så vi har behandlat det då. Att man säger. Jag gjorde till exempel en machine learning modell för att extrahera information från fakturor och med konsult information och sen föra in i ekonomisystem. Och ibland kunde det saknas kanske aa men något belopp eller datum eller sådär. Då beroende på hur dyrt felet är så kan du ju chansa. För det var inte speciellt dyrt att föra in något fel i ett ekonomisystem men om vi skulle till exempel ta att man ska beställa stora mängder leveranser i en supply chain kedja så skulle det kanske varit väldigt dyrt och vi hade skickat det direkt till manuell hantering. Så man får göra en bedömning av hur dyrt felet faktiskt skulle kunna vara liksom. Och hur mycket man kan chansa för att den här konsult AI modellen då, där körde vi mest på och blev det fel så kunde vi ändå korrigera det ganska och se. Det kom upp till ytan ändå men om dem här leveransgrejerna som vi också var inne lite på, när vi skulle beställa större mängder kött. Då var det såhär det får inte bli fel liksom, det får inte bli fel alltså för det kommer kosta såhär men runt 20-50 000 per fel i beställningen. Så att då är det såhär allting som saknas blir det manuell hantering på det liksom och då kommer vi inte behandla den datan. Jag vet inte R4 om du, har du?
10	R4	Nä jag tyckte du var ganska spot-on. Jag har nog inget riktigt att tillägga just kring det där så. Jag skulle säga att det är så vi hanterar i dem projekten jag har varit i än så länge.
11	FC	Ja, snyggt! Okej nästa fråga handlar om aspekten Data Accuracy och Data accuracy innebär då i den omfattning att data faktiskt är korrekt och pålitlig eller att den kanske till och med är certifierad. Så stöter ni på utmaningar med detta, med data som inte är korrekt eller pålitlig i ert arbete? Och hur skulle ni hantera det isåfall?
12	R3	Ja, det har vi väl. Om vi tar vårt senaste exempel då vi använde, ehm vad skulle jag säga, Machine learning modeller och cloudtjänster för att hämta hem väldigt mycket olika data från olika datakällor och visualisera det för att förenkla bearbetning av data. Och då blir det ju att du snuddar vid olika suspekta källor, vi hämtar hem från Google vi hämtar hem en del från aa men andra öppna datakällor och då får du ju friskriva dig från allt det och säga att datan rör sig i ett brett spann liksom och vi kan inte garantera att allting stämmer. Men om man skulle till exempel göra någon process för ett företag istället och använda sig av deras data istället. Då skulle man väl i ett annat skede säkerställa att all data verkligen stämde och sen kanske göra en prognos innan du påbörjar hela utvecklingen och säkerställa att allting stämmer men om du tar data från olika källor så får du väl friskriva dig från allt det där, att det kan vara fel liksom. Jag vet inte riktigt om det var ett svar på frågan men.
13	FC	Jo men typ så att det ni skulle göra egentligen då för att kontrollera om datan är korrekt är att ni börjar alltid med att visualisera den för att titta på den för att skapa en bättre förståelse då eller?
14	R3	Ja.
15	FC	För att se om där ligger några outliers alltså nåt som är sådär avvikande värden eller så då.
16	R3	Ja men exakt. Exakt.
17	FC	Okej. Ja men en annan aspekt då är Data Currency och det handlar då om hur aktuell datan är eller huruvida den är aktuell överhuvudtaget. Är det nånting som ni känner att ni stöter på i ert arbete att det är problem med att det är såhär inaktuell data? Eller är det så att ni tar ni hänsyn till tidsstämplar på data eller så överhuvudtaget?
18	R3	Alltså i de flesta fallen är det ju realtidsdata vi arbetar med att det är data som kommer in direkt och då har man aldrig behövt ta hänsyn till att den skulle vara gammal. Men nä jag kan faktiskt inte säga att datan jag skulle bearbeta har varit för gammal för det har varit mycket mer fokus på att hämta hem data som är liksom realtid då alltså något som precis har gjorts eller hämtats och göra predictions utifrån det. Jag vet inte om du kan komma på något R4?
19	R4	Nej, jag tänker i de fallen så har det inte varit något affärskritiskt att datan måste ha en viss nylighet i de fallen vi iallafall har jobbat med än så länge. Det är klart att det alltid är något du måste titta på att datan är så pass up to date att du kan använda den. Den kontrollen görs ju när man gör den första genomgången på rådatan tillsammans med den verksamheten som vi bygger modeller åt. Men jag skulle inte säga att det har varit ett problem så att vi behöver göra några åtgärder mot det, i de projekt som vi har gjort då.
20	FC	Ja nä men precis, det är ju det vi får utgå ifrån. Okej en annan aspekt är Data deduplication och det handlar då om att det inte ska finnas dubbletter att det bara får finnas en enda tupel med korrekta värden från entitet så att man då ersätter alla dubbletter där så att det bara finns en som

		refererar till en riktig entitet, att det inte får finnas några dubbletter eller så. Är det något problem ni stöter på jag tänker om ni hämtar in data som du sa från olika system eller så så skulle det ju kunna vara så att du får dubbletter på det viset utifrån olika system kanske.
21	R3	Ja jag har stött på det i olika sammanhang. Alltså antingen sorterar du ut det i ett ganska tidigt skede genom queries där du kan ta SELECT DISTINCT till exempel eller ja jag tror det är distinct som tar ut en av dem. Men sen också att jag när jag bearbetar datan så brukar jag föra register på liksom vad datan innehållit. Att jag kan ta all data som jag extraherar till exempel från ett dokument och sen den och lägga den någonstans och sen så jämför jag den hela tiden så att datan är unik liksom. Det finns ju olika skeden att undersöka om vi har behandlat den här datan liksom. Men det skulle till exempel kunna vara ett exempel hur man undviker det då, att inte bearbeta datan två gånger. Så att innan du gör en comparision mot någon databas där du har lagt in allt du har behandlat. Och det är också ett ganska bra sätt att upprätthålla traceability. Så att all data du har behandlat, så att till exempel om vi skulle utveckla en modell på ert företag och ni vill se hur mycket data vi har behandlat och liksom göra olika rapporter för ni kommer behöva det på era styrelsemöten för att se hur bra den har presterat. Då kan man lika gärna sätta upp en sån databas där du lagrar allting och sen kan du göra jämförelser om du har använt den innan.Jag tror det har varit det vanligaste sättet.
22	FC	Yes. Du har inget att tillägga R4? Eller något annat?
23	R4	Nej, men normalt sätt i dem projekten vi har gjort så har man ju i de fallen man har hämtat data ifrån olika källor som kan ha konflikterande information så måste man ju på något sätt använda någon slags unik identifierare för varje rad när man för samman datan till en gemensam databas. Så att det är oftast i det steget, normalt sätt så profilerar man datan först och tittar på hur datan ser ut och hur ser strukturen ut och sen så gör man det då att man korrekterar så att den håller rätt format och att den är up to date osv och sen när man för samman datan från flera databaser där det kan finnas duplicerad information så måste man ju använda någon slags identifierare som gör varje rad unik. Och det kan man använda eller det vi använder mycket då är Google Cloud platform så finns det verktyg för att göra det.
24	FC	Ja okej. Då går vi vidare till den sista aspekten då som vi har identifierat genom litteraturen och det är Data Consistency och det handlar om i vilken utsträckning data presenteras i samma format, ifall datan ni hämtar in är kompatibel med tidigare data. Om ni ser att det finns några särskilda utmaningar inom detta? Och i sådant fall hur ni skulle gått tillväga för att hantera det om ni skulle stöta på det?
25	R3	Ja, alltså det där har vi ju sett att det är svårt att göra om du ska bearbeta stora datamängder om dem inte innehåller samma variabler så kräver det en hel del utveckling för att göra datan enhetlig och få den i samma form liksom. Och då kan det ju röra sig om att vissa variabler saknas i vissa system eller att fakturor har helt olika layout inom olika parametrar så att det vi har försökt göra nu jag och R4 är att till exempel göra modellerna mer så att dem klarar av alla olika cases liksom. Innan har vi mest fokuserat på en lösning om dem innehåller dem här parametrarna men vi försöker bredda scopet och göra den så generell som möjligt så att om man till exempel arbetar med faktura att man gör att den kan klara alla olika designer och den kan tackla bort fler variabler om den innehåller onödig information så att man tränar den på det sättet istället då istället för att stirra blind på just dem här fem som vi behöver då. Så att det förekommer helt klart att det saknas mycket data men man får bara ta täckning och göra det så generellt som möjligt, så här insamlingen och bearbetningen.
26	R4	Eh kan du ta frågan en gång till så jag hörde rätt nu innan jag besvarar den?
27	FC	Nä absolut. Det var ju då Data Consistency så det handlar om i vilken utsträckning datan presenteras i samma format eh som då när du hämtar in som vi sa exempelvis från olika system eller att det kan vara att det är utformat på olika sätt att det då är kompatibelt med tidigare data alltså den nya inhämtade datan.
28	R4	Mmm precis eh det viktigaste där skulle väl jag säga och det har vi gjort för många av våra kunder är att man måste börja rent organisatoriskt och sätta standard i organisationen och bestämma okej så här vill vi att vår data ska se ut. Det är det första steget och sen därefter måste man ju se till att man när man i stabiliseringsprocessen av datan när man gör en profilering av datasetet att man kollar okej följer den här strukturen den strukturen vi har satt. Och gör den inte det får man använda olika typer av verktyg till exempel Google Cloud eller andra typer av visualiseringsverktyg för att formatera om den här datan. Jag kan gissa att det kan vara datum på fel format eller att man skriver stora eller små bokstäver eller hur man skriver ut en postkod. Det

		kan vara typiskt sådana saker som man stöter på har olika format i olika system som man behöver först definiera vilken standard man vill ha i sitt slutsystem och sen bör man formatera om varje respektives systems format till det här som man önskar då liksom.
29	FC	Ja, okej. Då ska vi se, ETL. Det är ett vanligt begrepp som står för Extract Transform Load och som används just när man hämtar data från olika system och samlar det i ett nytt. Och då är frågan där om ni använder er av ETL för att säkerställa datakvalitet inom Big Data Analytics om ni vet att det är något ni själva suttit med eller om det är någonting som görs innan ni får datan?
30	R3	Ehm. Jag vet ju att vi gör det vi har ju kompetenser inom F3 jag vet att det har varit ett hett begrepp inom ämnet såhär Data Management och våra kollegor har sysslat mycket med det här ETL. Sen har väl vi gjort det i lite i mindre skalor du och jag R4. Att vi har korrigerat data vi har laddat upp precis som vi gjorde du vet i Google Cloud miljö att vi korrigerade små så det blev i samma form. Men det är inte något jag har jättemycket kompetens inom. Men det används helt klart liksom på området såhär Data Management och hämta data.
31	EU	Okej så det är något som används för att förbättra datakvaliteten men ni gör inte det kanske?
32	R3	Ah nä oftast inte bara i vissa små fall om man vill ha det i samma form. Det kan vara ett vanligt use case. Och att du då i det steget dels definierar datamodellen du vill ha och sen applicerar olika typer av ETL-flöden för att få det format du vill ha. Vi gör inte det jättemycket men det finns vissa fall där man gör det av en lite lättare typ men då kan det handla om hur transformationen är, standardisering av formatet eller att du vill applicera en nyckel på datan och joina tabeller. Men vi gör inte det särskilt mycket men det görs kan man säga.
33	R3	Ett exempel är att vi skulle samla in väldigt mycket information om olika bolag för att hitta ett bolag som till exempel då är intressant att förvärva. Så om vi hämtade väldigt mycket bolagsinformation från en sida som till exempel Ratsit som är liksom såhär nyckeltal för bolag och sen från en annan sida. Och Ratsit skriver kanske inte ut Aktiebolag efter varje namn men det gör den andra sidan, så då kan man liksom putsa den här datan och matcha det så att alla dem här bolagen lägger ihop den här datan på samma gemensamma nämnare. Eller annat som till exempel bolagsnumret matchas med den och sen så bara putsar vi ihop den datan så kan vi få samma och få in den datan tillsammans då liksom. Så det är väl ungefär så jag och R4 har använt ETL då men jag vet att det finns väldigt mycket mer avancerad sån Extract Transform Load tekniker.
34	FC	Ja, amen det var väl ett jättebra exempel. Då kan du ta över där egentligen.
35	EU	Ja. Ehm sen utöver dem här dimensionerna då så har vi även identifierat lite andra utmaningar i litteraturen som vi också tänkte ställa lite frågor om. Den första är urval av data samples, och i litteraturen så står det att även fast företagen har stor kvantitet av data så innebär det inte att samplet som man använder sig av är tillräckligt representativt för hela datasetet då. Alltså större är inte bättre. Men hur gör ni använder ni er av data samples?
36	R3	R4 vill du svara på den? Det såg ut som att du nickade.
37	R4	Nä alltså jag skulle säga att vi använder inte, som när vi uttrycker modeller så använder vi ofta någon av molntjänsternas som Google eller Microsoft Azure och där finns inbyggd funktionalitet där vi gör all typ av olika statistisk sampling av datan. Så vi gör det lite mer, vi gör det men det är mer drag and drop och klicka i den processen.
38	FC	Så det är inte så att ni aktivt själva sitter och väljer ut data eller då gör någon och skapar en logik som plockar ut åt er, utan det finns redan?
39	R4	Oftast inte, utan man kan välja mellan olika typer av samplingsmodeller. Det finns ju vanliga som gör ren random samling eller man säger en variabel som man använder som riktlinje vid samplingen eller en viss fördelning men vi gör liksom den mer genom att klicka i en parameter i utvecklingsmiljön snarare än att man gör det rent. Vi sätter upp det så att det görs.
40	EU	Okej, men blir det aldrig så att liksom modellerna skiljer sig väldigt mycket beroende på vilket sample ni använder när ni bygger modellerna? Alltså att dem representerar olika saker?
41	R3	Eh det kan det göra men det beror ju också jättemycket på hur den underliggande datan ser ut. Jag har än så länge inte jobbat med något dataset som har varit så, alltså om du jobbar med ett väldigt litet data och den datan varierar väldigt mycket så kommer ju såklart samplingen påverka väldigt mycket hur modellen beter sig. Men jag tror alltså hittills så har det inte varit ett problem.

42	FC	Men hur tror du att ni skulle alltså det handlar ju just om att man vill ha en rättvis representation av datan där i sin helhet, men hur skulle ni hantera det om ni ser att det har plockats ut ett dåligt sample? Skulle ni då göra om det igen med samma mjukvara eller hur tror du att ni skulle hantera det?
43	R3	Eh, det var en bra fråga jag vet faktiskt inte om jag har något rakt svar på den. Det beror ju kanske lite på om vad som är problemet om det är så att det är väldigt lite data och vi då kanske måste utöka den på något sätt. Som ett första steg. Om det är så att även fast det är tillräckligt mycket data men det varierar så otroligt mycket och att det är det som påverkar samplet, så vet jag faktiskt inte. Då har jag inget riktigt bra svar på det.
44	FC	Nä men jag förstår i och med att ni själva inte har stött på det heller så är det ju svårt.
45	R3	Nä men precis vi har ju stött på det teoretiskt sett men kanske inte i praktiken.
46	EU	Yes, nästa utmaning då handlar om felmarginaler. Då det poängteras i litteraturen att det är viktigt att identifiera dessa felmarginalerna i datan för man menar att datakvaliteten aldrig kommer att bli perfekt och att det alltid kommer finnas vissa fel eller avvikelser men hur ställer ni er till att det kan finnas felmarginaler i den datan som används? Och hur hanterar ni dessa då?
47	R3	Ja alltså, när man gör olika predicitons så brukar man ju generera en procentuell säkerhet i hur accurate dem är. Om vi tar ett konkret exempel så om vi läser in era handskrivna föreläsningsanteckningar och jag ska digitalisera dem och skicka upp dem i ett system så att alla kan ta del av dem. Så kan man lägga in osäkerheten då var det gäller till exempel b, era b kanske ser ut som 6 och det vet jag då liksom. Så det kan dra ner den procentuella säkerheten hela tiden om jag gissar på om era b är b eller om era b är 6:or. Så då kanske det ordet som innehåller b kanske är 95% säkert för att det är antaganden man får göra så man får sätta då en gräns med vilken säkerhet man vill jobba i. Om vi då bara skulle vilja köra 100% eller 99% eller 98% så kanske vi hade fått ett lite sämre resultat så man får se lite och arbeta fram vilken threshold som faktiskt fungerar och göra tester med vad som skickades in och vad som kommer ut och sen kan man väl dra en slutsats med vilken threshold som passar just här.
48	EU	Ja, har du något att tillägga R4?
49	R4	Eeh, nä men tillbaka lite till det faller ju oftast lite på vad det är för typ av use case liksom i vissa fall kan man tillåta sig själv att ha ganska stora fel men i vissa fall måste det vara 99% rätt varje gång. Och då kommer man väl tillbaka till hur ser den grundläggande datakvaliteten ut och hur kan man påverka den för att få en högre procentuell eller bättre accuracy.
50	EU	Okej, men mäter ni den på något sätt då? Alltså accuracyn eller datakvaliteten?
51	R3	Alltså Accuracyn och datakvaliteten blir ju lite olika, det kan ju iförsig påverka Accuracyn.
52	R4	Vi pratar nog lite om olika saker alltså både du och jag R3 vi pratar ju om Accuracyn där vi konstruerar en machine learning modell vad den har för accuracy typ träningsmodell beroende på testdata vi förser den med exempelvis.
53	EU	Aa okej så den jämför egentligen då traindatan med testdata och hur accurate det är.
54	R4	Aa men exakt. Du tränar den på ett träningsset och sen så providear du ett nytt set som den inte har sett förut som har samma format och sen testar du och ser amen hur bra säkerhet har den när vi testar på det här datasetet istället.
55	EU	Så då mäter man egentligen modellen, hur bra kvalitet det är på modellen?
56	R4	Precis! Precis, så det är modellens träffsäkerhet.
57	EU	Ah okej, men mäter ni på något sätt datakvaliteten? Eller på något annat sätt?
59	R4	Eh ja, alltså i terms of det vi pratade om förut liksom completeness, format på data och så. Och det gör vi då genom att använda verktyg som Google och Microsoft och då finns det verktyg som kan ge oss den här statistiken då ganska rakt av. Du kan få upp en visualisering över till exempel en tabell som ligger i excel där du får information hur varje kolumn hur complete den här informationen är, hur mycket följer den ett visst format och hur många records saknad också kan man få statistik om det då och få reda på vad för åtgärder man kan göra för att förbättra det här.

60	EU	Okej, bara en liten fråga till angående det här. Det här med felmarginaler då om ni mäter ut att okej den här kan få vara 95% och den sen inte uppfyller det och det är en större felmarginal än så. Hur hanterar ni det?
61	R3	Vanligtvis, vi jobbar mycket med att tillämpa AI och Machine Learning i Business Processes så vanligtvis så ersätter ju våra modeller ett led av steg som verksamheten utför så det kan vara någon Accountant ute på något storbolag Scania liksom. Så det man gör är väl att ärendet går väl som den hade gjort innan vi kom dit och automatiserade den med våra modeller. Så om den är för osäker så får den gå till manuell hantering. Då kan det ju komma som förbättringspunkter att personen som fick det här ärendet från modellen då när han var för osäker, han kan det finns olika verktyg för honom att lära modellen som är liksom att så här ska du göra nästa gång och pekar ut och det här är inget b det är en 6:a. Och då kan man retraina modellen så förhoppningsvis kommer inte det tillbaka då liksom.
62	EU	Okej, nästa utmaning då handlar om datarengöring och det kan ju då användas för att lösa problem med saknade värden, syntaktiska fel eller syntaktiska skillnader mellan olika dataobjekt, eller då bara att det är ursprungligt låg kvalitet på datan, eller avsakning av metadata. Hur använder ni er av datarengöring?
63	R4	Amen det här var jag inne lite på när jag pratade om hur vi använder verktyg som Googles verktyg Data Prep som vi använder för det här. Som är just det att vi importerar in ett dataset och sen får man information om hur kvaliteten är på den datan som sagt hur väluppfyllda och kompletta dina records är eller hur bra dem håller ditt standardformat. Och sen kan du tillämpa alla olika typer av rensningsmetoder på det, färdigbyggda sådana metoder där du kan till en viss grad städa upp datan till det formatet du vill ha.
64	EU	Men är det då rå data när det kommer till er eller har det gått igenom en rengöringsprocess innan dess?
65	R4	Det beror helt på från kund till kund. Vissa kunder är duktiga på att hålla datan på ett bra format och då behöver man inte göra så mycket utan mer bara handlar om att man vill standardisera till ett format som du själv tycker är trevligt. Men i vissa fall kan du få dataset som är ganska stökiga, framför allt om det kommer från olika system i en verksamhet som inte har någon tydlig intern struktur för hur man arbetar med data. Man kanske får ha någon typ av masterdatahantering i verksamheten. Men då kan det ofta krävas att man behöver göra mycket sån städning av datan.
66	EU	Mm, men vid dem här tillfällena då hur gör ni för att behålla en korrekt representation av datan efter rengöringen? Så att innebörden av datan liksom inte ändras?
67	R4	Eeh, jadu det är en bra fråga. Jag tror väl att det kommer tillbaka till det att man innan man börjar rensa att man definierar vilket format som krävs för att göra det här uppdraget eller bygga den här modellen, så att man har en definierad standard innan man börjar rensa. Så så länge man förhåller sig till den standarden som är satt från början så kan man säkerställa att datan håller den som du utryckte det att den fortfarande representerar.
68	FC	Till exempel om det är så att det saknas värden så kan man ju fylla på med värden. Det är väl egentligen också ett sätt att rengöra, är det något som också skulle göras av dem här programmen också eller är det något som ni skulle sitta med själva så att man sitter och gör det manuellt?
69	R4	De programmen vi använder gör inte det här själva utan då instruerar man dem till att göra det. Det kan till exempel vara att man det beror lite på vad det är för parameter man fyller i. Vissa parametrar vill kanske bara ha ett värde och det spelar inte någon roll vad det är för värde och då kan du till exempel fylla upp ett NULL värde till en nolla, vilket är ett ganska basalt exempel.Men dem verktygen vi använder gör inte det utan då måste vi instruera dem till att göra det.
70	FC	Men hur känner ni där, är det bättre att få rengjord data eller hellre att få rå data som ni sen själva får städa upp för att då inte datan inte ska tappa sin betydelse. Alltså att den har blivit för modifierad, är det en utmaning och risk ni ser att datan blir för modifierad?
71	R4	Alltså i vår process då eller menar du innan vi får den?
72	FC	Eh egentligen generellt.
73	R4	Okej, eh jag tycker inte att det har varit en utmaning än så länge. Och som svar på den frågan så ser jag hellre att man får den data som är någorlunda rätt strukturerad så att man inte behöver sitta

		och göra allt arbete själv. Speciellt också kanske för att vi jobbar i en bransch där man får betalt per timme och då vill man lägga så mycket som möjligt på att bygga det som kanske egentligen skapar värde för våra kunder, snarare än att sitta och städa data. Men det är ju såklart också från fall till fall. I vissa fall behöver du kanske mer kontroll på datan så då kanske du behöver göra den städningen själv för det är till slut du som ska få in den i modellen.
74	EU	Ja, bara ett litet tillägg där också. det kan ju som du sa ta ganska mycket tid och kosta en del att genomföra den här datarengöringen? Hur ställer ni er till det?
75	FC	Är det en avvägning ni gör med kunden eller är det ni som tar beslut där när datan är ren nog? Eller hur ställer ni er till det?
76	R3	Ja, lite det R4 var inne på så vi blir ju betalda av kund för att utföra dem här tjänsterna så oftast har det varit så att kunden är ganska noggrann med att allting är färdigt när vi kommer dit. Så vi har initiala möten där vi ställer krav att det här och det här ska allting vara på plats för att det som kan hota deadlines och utvecklingsprocesser är ju typ accesser och data då, att den inte är på plats. Så vanligtvis brukar man ju ställa ganska tuffa krav gällande dem här frågorna och kunderna är ju med på det då och villiga att ta fram detta så snabbt som möjligt och då till projektstart. Så hittills har det varit att kunden har den här uppfattningen och har fixat så mycket som möjligt innan och sen när ni kommer hit så utför ni det som ni kan och som vi inte vi kan.
77	R4	Sen kan det ju vara dem fallen där det är så att kunden inte kan det här själv. Och då får man ju fundera på om man kanske ska inkludera det här i projektet jag menar att förberedandet av datan är ju 80% av arbetet och själva modellerna är det lilla om du har bra data. Oftast går det ju inte att göra om du inte har vettig data och ibland kanske vi behöver komma in och hjälpa till med den biten också för att dem kan inte göra det på egen hand.
78	EU	Okej, men då verkar det ju som att ofta är billigare att lägga ner dem här kostnaderna på datarengöringen antingen om det då är ni eller kunden som gör det?
79	R4	Aa, precis det är ju ett beslut varje kund får ta då beroende på vilken inhousekompetens dem här. Det är klart ofta om man anställer konsulter för det här så har man kanske inte så mycket inhousekompetens generellt sett. Så då kan det bli att vi får göra den biten också men då ser man ju hellre att vi gör den processen också så att man kan göra hela projektet ‘än att man får skitdata liksom.
80	R3	Jag tänkte flika in där lite att det vanligaste som jag ser just nu när vi tillämpar data science på kunders behov så rör det sig i de flesta fallen om ekonomi alltså att du ska tolka kvitton, fakturor, invoices och så rapporteringsgrejer och det kanske är därför vi har behandlat dem här frågorna. Men det brukar inte förekomma för inom ekonomi så är det redan ganska tuffa krav från finansinspektionen, från reviderings alltså så här, att det saknas data i ett dataset där det är kunder som fakturerar det brukar liksom inte hända, det är inget normalfall. Men om vi skulle tillämpa data science på mer skräddarsydd för varje individ på LUs universitet. Då hade det blivit jävligt mycket jobbigare att liksom få all data enhetlig och större arbete på den. Så jag tror att jag och R4 i dagsläget får ganska mycket av det här gratis på grund av att vi har borrat oss in i ekonomispåret där det redan är ganska uppstyrt så att säga.
81	R4	Ja datan är ofta strukturerad sen innan. Så det är något värt att ta med sig ändå.
82	EU	Yes intressant tillägg. Det är ju bra att ha lite olika perspektiv på det. Men yes, vi har en sista utmaning här då och det handlar om ledningens stöd och engagemang, att det är en utmaning då inom datakvalitet att få ledningen med sig. Känner ni att ni får bra stöd från ledningen när ni jobbar med att ha en god datakvalitet?
83	R3	Aa det blir ju speciellt nu i och med att vi jobbar som konsulter. Menar ni kundernas ledning då eller vår interna ledning?
84	EU	Alltså ja denna frågan blir ju lite svår till er som arbetar som konsulter. Vi tänker väl egentligen företagets egna ledning men kanske inte i detta fallet då.
85	R3	Nä, jag kan nog svara ändå alltså jag antar att det är lite mer mot kunds ledning. Men jag tror att många av dem här processerna är det här att ledningen vill tillämpa det här då det blir två grejer, dels att man städar upp och man rotar omkring och tvingar businessen att hålla data på bra och uppdaterade former. Så att mycket av dem här initiativen att digitalisera datahantering och managementsystem och sånt blir ju att man kommer med en stor piska på till exempel

		insamlingen av data så att allting är up to date. Och det kan ju vara mycket med det här nya GDPR såg vi, att det blev att nu när vi ändå är igång och rotar omkring och sätter datan på enhetlig form så kan vi lika bra dra igång massa med digitaliseringsinsiativ nu när datan är så lättillgänglig och sitter på samma form. Så vad jag har märkt så är det verkligen stort engagemang från ledningen att all data ska vara up to date liksom. Det är liksom en av deras hjärtefrågor liksom att det är ett high-en och all data är up to date för det kräver bra data för den här digitaliseringen.
86	R4	Ja, jag tror också att de kunder som kommer till oss och vill ha hjälp med detta har ofta kommit lite längre i sin digitaliseringsresa. Dem vet att vi kan tillämpa den här typen av tekniker i verksamheten och då har man kanske oftast en organisatorisk struktur som redan stöttar datahantering, informationshantering i verksamheten. Färdig liksom, dem har en governance-struktur kring det. Sen gör vi ju andra projekt också där vi hjälper med allra första steget att man sätter en struktur och har en person till exempel en CIO eller en CDO som har tydligt ansvar för datahanteringen i verksamheten. Så när jag och R3 kopplas in då så finns det ofta en sån struktur kanske på plats, alltså det finns ett färdigt tänk. Och det tror jag är jätteviktigt för att man ska kunna göra dem här projekten. Man måste börja kanske med någon organisationsstruktur och visa kanske att data har ett värde och är en resurs i verksamheten. Och även om man kommer ut till en kund som kanske inte ger något sånt stöd från verksamheten så är det jätteviktigt att ha för att kunna ta sig vidare i den här resan och använda den här typen av avancerade analystekniker i verksamheten. Att man först har satt den, alltså struktur och hantering av grunddata, masterdata och så vidare.
87	EU	Ja okej amen bra. Då tänkte jag bara höra om det finns några andra utmaningar som ni kommer att tänka på som ni känner att vi har missat?
88	R4	Ehm. Jag tror lite såhär. Tillgängligheten av datan är oftast klurigare än man tror. Det är ofta man kommer till kunden och tänker kan inte ni göra något coolt och avancerat här och sen så har dem kanske inte så mycket data eller den är rörigt strukturerad sen innan. Man har svårt att plocka ut någon data att ge oss och det känns som många tror att man kan komma dit och bygga något tufft på väldigt lite information. Så att bara att få ut själva datan tycker jag är en stor utmaning.
89	R3	Jag tycker en stor utmaning är att tackla GDPR. Alltså det förekommer ju ofta ett namn liksom och jag måste, det blir ju ytterligare en process. Att se till att all min bearbetning och processering av data är GDPR-compliant. Om jag rensar allt, eller om jag extraherar jättemycket data från olika dokument så är det ju ganska lätt att det kommer med massa namn där och personnummer och sådär. Så just den aspekten där tycker jag är lite jobbig, det har jag sett på mina usecase.
90	FC	Ja, okej intressant. Det har vi faktiskt inte tänkt på.
91	R3	Nej, den kan ni ju skriva med som någon liten hemlig slide kanske haha.
92	EU	Amen bra, jättebra svar. Jag stoppar inspelningen nu en sekund bara.

Bilaga 5 - Transkriberingsprotokoll T4

Verksamhet: F3 - Revision
Medverkande: Respondent 5 (R5), Respondent 6 (R6), Frida Carlsten (FC), Sofia Nyberg (SN) och Emelie Uddenäs (EU)
Datum: 2020-05-05 **Intervjulängd**: 71 minuter

Rad	Person	Information (Svar/Fråga)
1	SN	So if we start with, maybe you guys could tell us little bit about yourself like what is your role that you work with, how long have you worked with it and what kind of task do you have?
2	R5	Yeah, so I can go first. So I have been with F3 since 2012, but originally with the UK firm and I have purely been working with analytics manly to support audits so external but also a little bit of internal. Then I came to Sweden in 2017 on secondment to come in and help drive this analytics being done in the Swedish practise and then I liked Sweden so much I decided to stay permanently.
3	SN	That's nice!
4	R5	Yeah, so I have been here ever since. So I am a 100% analytics so that I would say that most of my time at the moment is spent supporting audit as an IT-specialist. So performing the more custom analytics for, on a particular client and then give the results back to the audit team so it just outlies things that are a bit unusual for them to follow up,
5	FC	Okay!
6	SN	Thank you!
7	R6	Yeah, so I also work in audit with F3 and my background is that I am a master student in Business Administration with accounting and I have no real… ehm I don't have any IT education or background so of sorts but kind of ended up in doing work for our analytics team within auditing. The task that we do is that we use Big Data or data sets from clients to perform our audit procedures, with the purpose of you know signing of the audit in the end. So I do analytics from another perspective, pretty much but there is special data quality questions in that as well so I hope that we can bring something.
8	FC	Yeah, sure! It is nice to have the different perspectives of it, so that is great. So should we start with the questions maybe?
9	SN	Yeah.
10	FC	So there are many different dimensions of data quality and the literature mentions a lot of different dimensions and we have chosen to study five* (six) different dimensions of data quality a little bit closer. And the first dimension that we are looking into is Information Completeness so that is if any data is missing or if it is broad enough for the task so to say, if it is sufficient for the task at hand. Do you have any challenges with that, with information completeness in that matter?
11	R5	Yes we can do, because particularly if the client is extracting the data. Sometimes the data we can, we can just tell that the data is incomplete just because there has been errors in the extraction. One of the ways you try to prove completeness of financial data is to try and reconcile it back to the general ledger. Because that sort of links the --- balance we see that as a good way of proving our data is complete, and sometimes when there is a significant balance difference between the data that we have and what has been recorded in the accounts then we know that we can't really proceed until that gets resolved. On non-financial data it is a bit more difficult. Say you are looking at

		something to do with supply masters and your list of, you got a list of supply masters with their name, their addresses, their telephone numbers, etc. That is a lot more difficult to try and prove if that list is complete or not because there is no real way, since it is a very static list, there is no real way to definitely prove that that is a 100% complete so that is a challenge there.
11	FC	Definitely, and how would you manage that challenge? So if data was missing what would you do?
12	R5	Ehm so with financial data… yeah we would query that back to the client, usually send them a couple of examples. So say for example we are reconciling a stock balance back to the stock account and we see like a mismatch we would send that example back and ask the client to investigate it. With the telephone, with the nonfinancial data side we will try and ask the client to provide a list, like a series of control tables where they have like a row count and we set at a certain point in time and we see based on if their row count agrees to our row count in the data that we have.
13	FC	Yeah okay, so in general you would go back to the client and ask for more specific data or yeah… to make sure it is correct.
14	R5	Yeah and that might lead… on the financial side that might lead to yeah have full reconstruction potentially. So yeah we… you know sometimes they have realised the mistakes being made. It is sometimes easier to get a full data set again and then try the whole process again.
15	FC	Yeah okay great. Do you have something you want to add to that R6?
16	R6	Ehh no I mean completeness and accuracy in the data sets are really essential questions when working with analytics in an audit and from an organisational perspective as well for the client and there is different ways that we can secure how the data is complete. It is either like R5 says, we compare different types of data sets so you get an extraction or do an extraction and you compare it to perhaps an aggregated report from the same system or another system to see that they match so that we know that we have a complete set. Or we can scrutinize the report logic directly, so kinda you know to see okay what parameters are used in this report, do they actually make a foundation that will get a complete report from this or is something missing. We might compare the total numbers of rows in the system towards the report as R5 said. So there is different ways to go about, that is a question from our side about efficiency. What will give us enough assurance to a reasonable amount of time. Of course the best way would probably be to test the controls within a system itself and the report logic, but that is often… it takes more time and it costs more money because you need an IT-specialist to do it. So aside to that, we usually call it either testing a control load information, which is testing the actual control that the client has in place to know that the data set is complete or we do an direct test, that is kinda you know either reconciling different reports or doing samples for reports. So let's say that this ehhh… yeah it is an invoice report and so we might collect the number of underlined actually invoices and reconcile it to the report to see that they are not missing.
17	FC	Mmm, okay great!
18	R5	I'm sorry, just one other point. Sometimes depending on what has been asked or what has been required if we… and we are trying to ascertain if the data is complete or not sometimes there will be an allowed sort of tolerance or difference, so that… or financial tolerance if it is within x amounts for example then that difference is classed as being quite trivial so it seems that the data is complete enough to proceed with the task even though it is not a 100% complete.
19	FC	Yeah okay. So moving on to the next dimension which is Data Accuracy, which means to what extent the data is accurately reliable and certified. So do you have any problems with data accuracy, meaning that the data would not be accurate or reliable, or that it needs some sort of certification?
20	R5	So I have worked in a project in the UK which was looking at customer data, it was for a airline, and they were trying to… the whole project was ascertaining how good the data was. It was a data quality exercise and what we did was that we took the data and then we fed that into a, it is like an interactive dashboard so the client could see how good the data was. And that particular example though is accuracy issues because a lot of the data stands from the customers inputting the information into the airline. You had for example first names were the first name was only one letter for example. Optional fields, if they did not have to put down telephone number for example, that was left blank. So you then start to getting, when that whole exercise was about building a customer profile, you then have issues with the debate that people could have mistyped their name

		for example and then you get problems like that. And then it is the same in I suppose in when recording presented transactions when people are logging like accounting postings for example. They could you know leave, if they could leave the nature of the transaction blank for example so you know it could potentially get you know misclassified for example.
21	FC	Yeah definitely good examples. And if you… How did you handle that problem that you hade with the data quality task that you did for the airline? How did you manage the challenge with the human factor, if you had that in mind for when the customers entered their data?
22	R5	Ehh yeah so that is a good question. So we used… Because we had access to the raw data we could pull out specific examples, we could adapt. The client could then go back and recontact customers if they wanted to. We also, I suppose depending on what they were specifically looking for, we could refine the analytics so we didn't treat… we could treat a blank entry as not being inaccurate data as such is was just missing. We basically gave… could give examples back to the clients for them to follow up.
23	FC	Okay and then they could choose if they wanted to follow up or not?
24	R5	Yeah, because I think they had… they have an obligation to make sure the data they store about people are accurate and up to date so if they didn't think the data was accurate they would go back and try and contact those people.
25	FC	Okay great. Do you have anything to add R6 or are you happy with the answer from R5 as well for your part?
26	R6	Yeah I am happy with that, it was good. I think essentially it is the same way how we work with data accuracy is very much the same perhaps as how we work with completeness in some sense. Perhaps with a an meaning that we might check the data directly or check the report logics and how the data is retrieved. But other from that, as R5 said, we might do some you know reasonable checks. I mean that you do kinda analysis of the data finding the outliers and returning them to the client If you go that way or checking them within in the system if you can do that to double check that they are actually correct from within the system. So there is nothing wrong in the actual transfer of the data from the database to the yeah whatever you are using.
27	FC	Okay perfect. So then the next question is about the dimension that is called Data Currency and data currency is about whether or not data is current and up to date. Do you see that you have met any challenges with up to date data? That it is not current enough or that you can actually use it, that it is usable.
28	R6	I think it is that the data is actually up to date is less of a problem in audit since we do, since we do our work sequential so it is kinda one year at a time or one period at a time. You kinda notice if you get a data set that is not for the period you are currently analysing or is for a previous period. You would notice thar within your completeness and accuracy checks. I think the data up to date… I have not really faced any issues with that. What about you R5?
29	R5	Yeah so I would agree with everything that R6 just said about it all. I have a couple of examples, so we… I was working on a forensics project looking, trying to identify potentially fraudulent transactions and payments made to suppliers. In a supply masters data when you have… you can sometimes see that it is multiple suppliers that have very similar names or very similar deviations which suggest that the data might not be current or up to date. Most of the times if there is a huge lead like a valid from or a valid to date field in the data, so that can usually help trying to track down when the data is most current. Sometimes we look at if a data system has exchange rate tables which people are updating we can see when people are changing the exchange rates that are used if there is multiple currency transactions coming in and we can look at that table to see how, when it was last updated, when asserting currency was last updated and by who.
30	FC	Yeah okay and when you...
31	R6	I can add that… I mean that is a good example, I have several clients actually that you know they have internal CRM-systems work for either the financial treasure operations from which they take exchange rates but at the same time they have external kind of financial instruments that are, which value is based on the banks exchange rates and of course they have differences in their accounting just because their data is not enough up to date. But then I guess for them it is a question about okay how much is the difference, is it reasonable? And then they have to manually kind of check and see if they need to correct it themselves.

32	FC	Okay, so then it is a manually correction of it?
33	R6	Yeah for them often.
34	FC	Yeah okay, so the next dimension I want to ask you about is Data Deduplication and it is essentially that you should not have two entities of the same in your database or your data storage. So how do you… Have you faced any challenges with duplications and how would you go a head to manage it?
35	R5	So I haven't really experienced an example of this, like most entities are being duplicated. What I have experienced is a raw data set, wetter it be financial or example… just for the example where there have been duplicate lines, data lines in the same data set which… where every single field is the same. That usually an extraction error, but depending on the level of the duplication we can sometimes fix that ourselves just using coding to group the lines together into one.
36	FC	Yeah okay. Have you encountered any problems with duplications R6?
37	R6	Yeah sure, I mean it is a common problem when we do our own analysis, at least for me. I guess it is a… the base of the problem is like R5 said is probably that you get a raw data set which lowest... where the lowest common denominator is not unique on every row so when you kind of aggregate the data set you get a lot of duplicates. Because when you join tables it will duplicate the number of rows because there is essentially two of the same in one of the tables and not the other So then as R5 said, then we usually kind of have to work around that with either trying to aggregate it into one row or kind of separate them by giving them additional columns which are unique then for each row. So that can cause kind of a lot of issues in the analysis but it is often… I would say that it is often… You kind of notice this when doing the completeness and accuracy checks you can notice when there are duplicate rows and also if you think that you have a lowest column of denominator you can always count the number of rows that had the same values put in that field to see if you have duplicates or not. So there is different ways on how… but it is also you know it is often a… in our case it is often also a manual kind of check doing an analysis of that.
38	R5	Yeah it is just to… yes it just to answer that. Yeah I think that there is two causes of duplicate data. It is the fact that there has been an error in the extraction or the data is genuinely how it is but we just don't have a unique identifier in the data. Say for example like a transaction number, a journal number for example. So we always, when requesting data we always try and ask for a unique ID or as R6 said we can depending on what fields have been provided we can create one ourselves by stapling the number of different fields together to make each line true unique.Something that we do have to consider because when we are handing over individual examples of samples to the audit team, sometimes they want to go back in and investigate the examples from the client. So they need to be able to find that specific transaction for example that we have identified.
39	FC	Okay I see. So the last dimension then is Data Consistency and that is whether the data is presented in the same format as previous data and that is compatible with previous data. Have you faced any issues or challenges within that area?
40	R6	I mean, our biggest data consistency issue is that we work with a lot of different clients that have different systems. And we run our standardized models or advanced reliability the way we always do, we cannot tailor them to fit the dataset that is provided by each client. So yeah that can be an issue that requires a lot of reworking from our side annually. And beside from that, I guess that when clients swaps systems or alters their report logics yeah it can be a data consistency issue that requires a rework of our work to make the Big Data Analysis to work.
41	R5	I would agree with what R6 just said. Sometimes even with the same clients on year and year basis even some extracted the data one year, and either that person moved on or it's been passed to a different person in that organization. I have for example experienced where the Data Warehouse looked different year on year so we don't have to adapt our analytics and our coding to feed in the data in order to manage that. Also, yesterday I had an example where it was transaction data for a client that was provided by two different people and one dataset had more field in it that the other. It was two sets of the same transaction. They had to be combined to one by basically just using some analytics coding to combine them together.
42	FC	Okay great example, thank you. So we're done with the dimensions of data quality. So now we just want to touch the subject of ETL. So ETL stands for Extract, Transform and Load. Is it something that you are aware of that you are using?

43	R6	Never heard of ETL. Is it a program or just an expression?
44	R5	It's a process. It's basically either we or they obtain the data extracting it or the client are giving it to us. And then we would process the data say for example SQL or Qlicksense and then we would load the data into our frontend, dashboard for example. So yeah we do have experience of that process.
45	FC	Yeah exactly. What challenges would you say ETL helps you with?
46	R5	I suppose with the E that obviously, sometimes we go out and extract the data directly from the client or sometimes the client provides the data to us. So the extract-process can a little bit different. But the actual process I suppose we're able to test 100% of the data population which helps from an audit perspective help enhance the quality we can claimed that we have analysed and profiled and sampled a 100% of the population. And sometimes the raw data of itself doesn't come in a very friendly fall out so it might be for example delimited data so there not in nice columns that you see in an Excel spreadsheet but like a pipe that used to separate each columns. So by transforming the data we help a non data specialist, it makes it more readable by transforming the data. And loading the data into something that is more accessible to a front end user. Take for example a dashboard that people can interact with. The end user gets more insights from their data.
47	FC	Great thank you. Then I will let you ask some questions as well.
48	EU	We have also identified a few other challenges from the literature. And we're gonna ask some specific questions about them. And we can start with the selection of data samples. The literature says that even though you have big quantities of data, this does not mean that the sample you're using is representative for the entire population. So we were wondering if you use data samples and do you think it's true that sometimes it is not representative?
49	R5	So if I go first and then R6 being an auditor as well will have a lot more of input in this question. So I believe sometimes we are asked to do riskbased sampling so we'll subtract the data through a serial of tests say for example an user who posed a lot of small values transactives or user who posed a lot of small times in the year or a posted transaction that being posted for an x amount for example. And those samples cause their more interested in the potential outliers and potential errors and potential cases of fraud they use drive the samples. It means that the samples are less focused on taking random samples and more driven to, the more interesting data lines. R6 will probably have a lot more to add to this.
50	R6	Eh..I dont know if I have so much. Sorry, I dont really know if I undersatnd the question, is it if we extract a portion of data that is big but not the full population?
51	EU.	Yes.
52	R6	Yeah okay so it's kind of a statistical question.
53	EU	Yeah like how can you make it represent the whole dataset and is it any challenge.
54	R6	Good question. Okay so either we do a full extraction and of the dataset so that we know that we have the full population and that it does represents the population. Or otherwise we can do a random extraction. Sure, it's pretty rare but we can do that. Of course that would kind of depend on how certain statistical criterias to say that it's representative for the full population. And that is why I think we do it pretty rarely and we only do it for a dataset that we know that a random sample would be enough representative for the full population that a statistical representation would be one of those bell curves thats evenly distributed over the normeless. But otherwise I would kind of say that we analyse the population and what it consists of etc and do an extraction of you know a stratified population type in the dataset that we think represents a certain type of transaction within the population but not the entire population itself. So it could be for example, typical thing we do is check accounts receivables for a client to know that they exists and get pay by their customers. And then we might do an extraction of only transactions that are referring to receivables that are very old, customers that haven't paid for 90 days. And that wont be representative for the full population but then we would have a riskbased population to analyse what we think is the more risk receivables. We have kind of standardized procedures to work with the statistics, how we work with statistics here. I would say the most common thing we do when we work with analytics is that we do the full population otherwise... it can of the base also to identify if there are derivations within the population that should be stratified you can have to know the struct of the population itself.

55	R5	Yeah it's our policy to, if we are performing Big Data Analytics, it's our recee to test the full population. Cause also if you don't test the full population you can't prove that the data is complete and accurate.
56	EU	But isn't that also a challenge, if you have a big amount of data, doesn't it take a lot of time to process the data and build the models?
57	R5	Ehm yes particularly getting the data can be a challenge. We trying the data in a text file to the admitted format. We try to keep it small as possible. So we have an online secure file transfer system that clients can use. Sometimes clients wants us to go and pick it up and transfer it securely on to our datasystem while we're there. Our tools that we use like SQL and QlickSense. SQL is like a database tool which allows you to lay your data into tables, and that can process billions of line of data. And it can take time. And our dashboards tools that we use QlickSense and Tableu can for example take a large volume of data. The processing time is sometimes quite long time. But that it is not so much we can do about that.
58	R6	I don't know what you say but I think the types of datasets that we can work with both of visualization and it is usually up to a couple hundred million of rows at least. You know, in some cases I have clients who have, online gaming companies, that have billions of transactions eh. For us is not feasible to do data extraction then for the client. I guess it also would be interesting to talk to them also and see how they do it. Because of course they must have full control of the data but there's a lot of small transactions. But yeah it takes time, it can be quite the process to handle large datasets.
59	EU	Yeah okay thank you. So another challenge that we have identified from the literature is the margin of error. Because the literature say that you can not get your data quality perfect and there are always gonna be a margin of error. What do you think of these error and how do you manage these margin of errors?
60	R5	Yeah so I think there will always be an element margin of errors. It's very rare to get a dataset that reconcile 100% complete for example. I think the mace case is there's an agreed tolerance you have I think the CTT that stands for Clearial Trival Thrushound which is where the difference is below a certain amount then it's of very low value so that's an acceptable margin of error. And then you have materiality as well which is to do with the financial value of the organization, I think. That sometimes you determine what an acceptable margin is.
61	EU	So you are aware that there some error, do you do anything to handle it, document it or something?
62	R5	Ehm if it's an error that is caused by extraction we would try and get a real extraction because that is caused by a human error, rather than with the margin of error and data itself. If it is with the data itself, as part of the documentation and reporting we do, we would show that we attempted to reconcile the data in an acceptable process. For example reconciling it back to the general legend. We would provide some real, what account reconcile and what account did not reconcile. Somewhere that margin of error would be noticed.
63	R6	I mean other than that, it's like with the accuracy of data, kind of… if it's not an financial error that we can explain with accounting terminology, lets say its registration date of the customer which can be pretty any date. You could analyse the whole dataset with a dashboard and you can investigate some what the difference is to find out what are these deviation based on and why do they occur. And if you can explain reasonable you can kind of exclude them or accept them as they are. If there's an input error oin the source of the data there's not much you could do besides you know try to look besides those error within the dataset.
64	R5	So we also do projects like we call advigering projects that's when the clients engage to tackle a specific problem. We go directly to the client rather than our audit team. We will pob limitations in caviax in the reporting if we need to. And then we would of course show them our reconciliation ourselves. Maybe going of a bit vintangeged, sometimes in our analytics we been asked to pull a series of transactions that meet these criteria and we get a whole of them that look like they're genuine transactions and there is nothing wrong with them. We call those False Positives. Cause they have met criteria but they are not of interests or they're perfectly normal transactions. So either they will be doubling some elements reporting just to say that they have been recognized as typical normal transactions or we will refine our analytics to exclude those transactions as with the analysis, trying to get them more refined.
65	EU	But you mentioned in the beginning that it has to been under a certain percentage, and how do you

		know it's under that, do you like measure your data quality in some way?
66	R5	So the cedermeteriality and the clearial thushour will be determined by the audit team and then if our data is out by a sense of map we will then consult with the audit team who will then consult to see how close that is around the itermetality or CTT. And they will decide if that is an acceptable tolerance or not. If it is we will proceed with the analysis, if it's not then we would have to go back and ask the client and follow to try to prove the difference downs so it's at a more acceptable tolerance for the audit team.
67	R6	I mean these types of materiality and CTT that is solely audit kind of terminology. We use it in data analytics as well but it has nothing to do with data analytics. It's a numerical value that we state which we find is a tolerable misstatement for a companies financial reports without them being so misstated that they do not represent the economical standpoint of the company. But we use it in data analytics as well because as you said, sometimes you can't get it fully right but we have to get it enough right to able to rely on the data. So we use it there as well.
68	EU	Okay. Moving on to the next challenge and that is within data cleaning or data mining and the representation av data. You can data cleaning to solve problems with missing values, syntactical errors or if it's just low quality of data. Or maybe if there's no metadata. How do you use data cleaning and what challenges do you think there is with data cleaning?
69	R5	Eh yeah this is an area that I know very well. Ehm so yes sometimes the data is very messy and so some typical examples are that we have dates that are split so half of the dates are american and half of the dates are non-american format. Just because the data have been extracted and put in to excel and excel automatically converted some of the data from american to the other way around. Some data cleaning do that, it trying to fix it by itself. That one with the dates, if we're using the dates as key to our analysis, we would probably play it safe and go back to the client and try to investigate and see if we can get an attempt to reextrac.. Sometimes our tools for example process or find financial values with a dot as a decimal, but a lot of swedish values have a comma instead. So we will use tools ourself and we can use excel to try to do that or as part of loading it in to our analytical tool softwares as we working on transforming and processing the data we can write code to correct those issues as we go on demand. And sometimes the data is sceemed? so we get a lot of datasets which are called delimited datasets and that's when a certain symbol or character is used as a column separator rather than having a table of data. So if the column separator is a comma for example and then in a an address line you brought a comma, that's what sceemes the data along by column, but let me know if this doesn't make any sense. We sometimes get the data out in the wrong columns for example and we can normally fix that ourself manually.
70	EU	Okay but how do you assure that you keeping the right representation of data? Like after the data cleaning how can you assure that it still represent the same data and that it's not manipulated in some way?
71	R5	So we would never change the underlying raw data. We would only change the data cosmetically to make it easier to process. And if we can't guarantee for examples that cleaning the dates for example we don't know that is 100% accurate, then we won't rely on that column. Either we would carry on with another or if we need this not valid column for this analysis we would have to go back to the client and we have like a quality procedure where the work of somebody performing analytics is always reviewed and that person would always check if the raw and underlying data has never been accidently manipulated as part of the process.
72	FC	Okay so that's how you assure that this challenge wouldn't appear, like it wouldn't be an issue for you? You don't manipulate the data in that sort of way.
73	R5	Yeah we would never have any reasons to change the underlying and raw data as part of our analysis. We can add new columns in the data and add things by ourself, but the data that we receive we never manipulate the underlying content.
74	EU	Okay and all the time and cost associated with data cleaning, how do you manage that? Is that something the company is prioritizing?
75	R5	So we I mean our analysis we're doing or we would never get a project where the purpose of the project would just be to clean the data. Cleaning the data would only be a part of the process of the analysis. But whenever a job involving analytics is costed, loading and transforming the data is a part of that costing and a number of hours is allocated into the budget to allow time to clean data. Sometimes the data comes in very good quality so hardly any time at all is needed to clean it,

		sometimes it comes in very bad and it's very messy. I think it largely depends on which system the data has been extracted from. If it's a small and old system the chances are more likely that the data is going to be very messy but if you take it from a big and well known system for exemple SAP just as an example, the data on the hold tends to be more cleaned and less cleaning is required.
76	FC	Yes, do you have anything to add on that matter R6? I don't know if you have been doing any data cleaning?
77	R6	Eh yeah I mean data can as usually unless it's the case that the data comes in a perfect manner, eh data cleaning is usually a part of the start-up process before you can..
78	R5	Eh we lost you R6
79	R6	Yeah I'm back. Ah sorry it shuts down randomly. But yeah it's a base kind of thing we do in the beginning of the analysis is to clean the data. It can be very time consuming and I mean to check whether we have actually after we have cleaned it if it is altered in any unwanted way and so on it's typically the same procedures as when checking that the originally data is complete. When you have your clean dataset, does it still reconcile and with the sense of the raw data does it still contain the same amount of rows and columns that you want. When you look at the data transformed, is it understandable, does the columns contain what you want them to contain in the same format etc. So that's the typical procedures that we do. But it's the same there you know, you can never have 100% assurance or confidence that you haven't altered anything if you have a very large dataset say 10 million rows, you can't check every row. At some point you can satisfy that it is accurate enough after it being transformed.
80	FC	Okay, and how would you check that? Would you check that manually? Let's say if you have 10 million rows maybe you won't go through every row individually or do you have an automatic process for checking that?
81	R6	Eh I would say there's a lot of manual checks involved and some of them can be automatic for example when you checking the total number of rows you can just write a short script to count the number of rows and compare it with the raw file. Otherwise it could be after you have transformed the data you sort it up in a table and you just sort it for example a date column should only contain dates and if you sort it you can notice if there is any rows that is not dates just like you did i sort i excel sheets or whatever. So it's those kind of manual checks that you do.
82	R5	Yeah you can write a script for example to pull out a list of all the unique dates for exemple so we can get a list and see if there is any non-dates as R6 said if that is an issue. For example we in a financial dataset it's expected that every line have some sort of financial value, we can write a script to pull out any lines where the financial value is blank and then that usually indicates that there is an issue with these lines. And if the dataset is small it's usually more efficient to do the checks in Excel and if the dataset is under like a million lines three lines in excel filtering we can usually spot if there is any issue with the data.
83	EU	Yeah okay. So we have one more challenge left and that's the challenge with getting the management support in the process of getting good data quality. Do you feel like there is any issues in getting the support from the management or investments from the management?
84	R5	Ehm, that's a good question. So I think there is two streams to that question. I think firstly there is a challenge of obtaining and extracting the data and sometimes there is resistance particularly if the management don't understand what we want to do. Or management have full sign-on but then IT who are doing the extraction don't understand why we need that data so usually in those data obtaining and extracting conversations we usually involve both somebody from management and finance and somebody from IT as we found that works better and then we have the sign-on from both groups. To any terms of the actual data cleaning and quality and the management's sight on that I think it really depends on the actual client because sometimes the dataset is very or the system is very old for example if maybe it's a smaller company they just don't have the time and resources to fix the data or investing in a completely new data extern in order to fix the data. Other times depending on the nature of the engagement the client is a lot more interested because if they are using a lot of personal data and with things like GDPR they have responsibilities to make sure the data is clean and accurate. So I suppose it I think with financial data sometimes yeah if they know that things are misposted they will go back and correct the pastings particularly if there is a large amount. So it is quite varied depending on the client I would say to summarize.
85	EU	Okay do you feel like they are more supporting in the process of doing the actual analytics from it

		or more invested in the other processes like cleaning?
86	R5	Yeah if we sometimes the people don't really know what they are doing because they are a small client or they haven't got the intracks yet sometimes they just don't know their underlying data so sometimes we can by analysis show them a very valuable insight if we for example I did a project in UK when we did a payroll analysis and looking to make sure you know so people were paid accurately and there were not any unusual payments. But we found in the data we received we found seven employees all registered to one property which they were not aware of at all and when they find something big like that they don't go back and investigate. If you find something really interesting and powerful they will obtain notice and they will support that because they are interested and because sometimes it get presented to the highest levels of management so I think they are supportive of analytics if we can convince them what they are going to get out of it in the end and like show it in a lot of visuals that we can put in the report where they can find really interesting insight or we can confirm that we help them identify or give them the assurance that there is errors in transactions to the suppliers for example. It depends on the client because some clients are really in to it and others can be a bit more resistant.
87	EU	Okay, so that was the last challenge we have. Now we just wondering if there are any other challenges you are managing or do you think we have missed out on anything?
88	R6	I think it was very good questions so I don't know if I have anything to add. Sometimes the data integrity and all this transformation those are new and usual challenges and sometimes you're past those points and then there can also be the challenge with make the actual analysis that you're going to do with the data purposeful. And if I have extracted a hundred million rows I want to show something that those low level detail extraction show us that we can't show on an aggregation report on ten rows. Then we have waisted both our time and our clients time, so sometimes it's the challenge to make a purposeful analysis of a huge dataset that's to the point with the issue that you are addressing. That's the final stage of Big Data analytics is to analyse and visualize and do the actual analysis of this and the pre work usually takes the most time but I guess it's the most fun part of data analytics to analyse it and it requires thought from the person doing the analysis and understanding of what we are doing so sometimes it is an issue for us in terms of finding the competence of doing this. You both need the IT competence, which I don't have for example, and you also need the understanding of accounting and auditing to make a purposeful analysis and some cases when IT persons do the whole work they miss the point and when financial people do the work they get the data wrong instead.
89	FC	Yeah I guess it's definitely important to have the right knowledge about the data to truly understand it and to be able to create a good analysis from it.
90	R6	Yes of course. It is essential.
91	R5	One thing I would add is that sometimes the system that we get the data from the data is stored across a lot of separate tables so we get several datasets that we might need to join together. And sometimes there's challenges to truly understand how the data links together and also joining the data together. Usually we try to join them on common fields but if you don't join them accurately you start creating duplicates in the dataset and it will start going wrong. So I think there is that challenge that we are not entirely sure how the data links together or we have to work out say five different datasets together, then we have a big risk of making the data quality not remain as good as before we started combining the datasets together.
92	R6	That's a good point I mean that's also in terms of the mainside of data quality this is a point when you have several different systems it is difficult for us and it is also difficult for the clients to understand as well. If they have a data system from which they store data in datahouse A and then use an API to transform the data and then store the data in datahouse B. and then we extract the data from datahouse B it is not sufficient for us to only control that the data we get is reconciling with datahouse B because the data comes all the way from the main system and there are two transferings of data between that and you know datahouse B. So a lot of clients have insufficient or no understanding and control over the whole flow. So that of course from an audit perspective that poses an issue because we have to explain why we find the data inaccurate and also from an operational perspective. It is of value for the client to know that the data reconciles all the way.
93	FC	So what would you do in that case? Would you ask them to extract the data from warehouse A?
94	R6	Yeah we could do it from the source or we could do constellations kind of you know reconcile datahouse B to datahouse A or we could test if the client have controls or transfers between the

		systems we could test them all the way fro datahouse B and do our work from there. But sometimes you know there are no controls and you have to go all the way to the source essentially.
95	FC	Oky that's good points. Thank you very much.
96	EU	Yes thank you very much, I will just turn off the recording.

Bilaga 6 - Transkriberingsprotokoll T5

Verksamhet: F4 - Revision
Medverkande: Respondent 7 (R7), Frida Carlsten (FC), Sofia Nyberg (SN) och Emelie Uddenäs (EU)
Datum: 2020-05-05 **Intervjulängd**: 42 minuter

Rad	Person	Information (Svar/Fråga)
1	FC	Då börjar vi att fråga om din roll som arbetar du som idag?
2	R7	Idag jobbar jag som något som kallas Senior Associate på avdelningen Advisory. Och Advisory har flera olika delar. På ena sidan är det Transactional Advisory och på andra är det Operations Strategic Advisory, och jag är då på Operations Strategic Advisory, OSA som det kallas. Och under OSA finns det flera tjänsteområden men där jobbar jag på under en avdelning som heter Governance Risk and Compliance. Mitt primära jobb kan man säga är att vi fungerar som specialister och stödfunktion till vår externrevision. Jag gör lite mer komplexa och datadrivna analyser på liksom våra externa revisionskunder. Det kan vara allt ifrån att man vill granska intäkter och kolla på alla fakturor på alla bolag och kolla att de är korrekta bokförda. Man vill kanske kolla på löner, har det blivit bokfört korrekt, hur ser könsfördelningen ut. Det finns en hel del olika typer analyser som vi gör, men näst intill alla analyser är tunga datadrivna analyser med många observationer. Och jag tror det är därför Jessica vidarebefordrade ert mail till mig.
3	FC	Jag förstår. Så du arbetar med dataanalysen då?
4	R7	Mm precis. Det är det primärt. Sen är jag involverad i en del interna projekt kopplade till automatisering och en mer datadriven verksamhet och ah att beslut ska fattas mer datadrivet istället för "gut-feeling". Man kan säga att 90-80% till externa kunder, sen 20% internt.
5	FC	Okej. Hur länge har du arbetat i den rollen?
6	R7	På F4 har jag arbetet i två år. Och just den här rollen har jag arbetat i som, ah man kan säga två år, men att jag har gått från Junior till Senior. Men två år på F4 med liknande arbetsuppgifter.
7	SN	Okej. Då går vi in lite mera på ämnet då. Och då har vi identifierat olika dimensioner inom datakvalitet. Som jag kommer gå igenom en i taget och fråga om du tycker att det finns några utmaningar inom dessa och i såna fall hur ni hanterar dem.
8	R7	Absolut.
9	SN	Den första är Information Completeness. Det vill säga i vilken utsträckning data inte saknas och har tillräckligt bredd och djup för att klara av den aktuella uppgiften. Ser du att det finns några utmaningar med de här och hur hanterar ni dem i såna fall?
10	R7	Ja asså det är ju något som vi granskar hos våra externa revisionskunder så jag tror mina svar kommer att bli ganska mycket, eftersom vi jobbar en del med de här utåt sett tror jag att jag kommer att hantera frågorna på ett sådant sätt, hur vi granskar de hos våra externa revisionskunder. Ah det är ju också så vi granskar de internt här då men liksom de här är ju en väldigt central del i vårt arbete när man jobbar med just revision. Att just säkerhetsställa att all data är på plats. Så det här är egentligen det första steget som vi gör när vi får in en stor datamängd. Om man tar som ett exempel så gör vi mycket, ah vi kallar det fullständighetstest, men det är ju egentligen Completeness test, utav våra externa revisioners hela bokföring. Det man gör då är att man tar in hur deras balans- och resultatrapporter ser ut och sen tittar på alla transaktioner och ser att alla transaktioner blir deras slutgiltiga balans- och resultatrapport och den rapport som publiceras som

		årsredovisning och som investerare tar beslut på. Så det är en väldigt central del att säkerhetsställa att de siffrorna är korrekta. Och den problematiken som ofta finns är att personer som jobbar på ekonomiavdelningar generellt sätt är ofta ganska, har ganska svag kunskap i hur olika exporter fungerar av data. Som ett exempel så är det för oss, det spelar egentligen inte så stor roll hur datasetet ser ut så länge man får med allting. Och att den som exporterar datat inte gör några filtreringar eller ta bort transaktioner. Och det är ofta något som återkommer att eh vi ber om ett fullständigt data, men personerna som man frågar, förstår inte vad ett fullständigt dataset innebär. Det kan vara fullständig för dem i deras mening, men för oss så vill vi ju verkligen ha allting. Och det är där som är den stora svårigheten eller utmaning är, att det finns en generell okunskap hos våra kunder när det gäller att hantera en stor mängd datamängder.
11	SN	Men om ni får data, och det då saknas data, hur hanterar ni det då?
12	R7	Ja men det är bra. Den första åtgärden, eftersom vi verkar i en tidspressad bransch och prispressad, så ber man ju då kunden oftast, man pekar på de felen man har sett i de kontrollerna vi utför. Till exempel, det finns flera kontrollfunktioner man kan göra på en huvudbok, men då sänder vi över de felen och förklarar vilka typer av transaktioner som saknas. Och ställer kontrollfrågor i, vilket datumspann är det ni har tagit ut de här transaktionerna, och hjälper dem lite påvägen. Eh och sen ber om man om ett nytt data. Skulle det vara fel också så får man göra en lite mer djupdykning i är det något handhavande fel hos den här som jobbar på en ekonomiavdelning eller är det något fel på systemet, varför presenteras inte alla transaktioner i den är typen av export. Vi hanterar det som så att vi tillsammans med kunden går igenom exporten utav för att få en bild av vad problemet kan vara. Och oftast så är det väl just olika typer av inställningar i exportfunktionerna som bidrar till de här felen då.
13	SN	Super. Då går vi över till nästa dimension och det är då Data Accuarcy det vill säga till den omfattningen datan är korrekt, pålitlig och certifierad. Ser du några utmaningar med den dimensionen?
14	R7	Ja asså utmaningen är ju den att många av våra kunder litar så starkt på sina system att de förstår inte att det kan bli fel och att data är felaktig. Och de är ju en ganska stor utmaning eller svår utmaning att tackla, men det vi erbjuder eller det som är ett krav också när man gör en revision är att det finns någon form utav det kallas ITGC, generella IT-kontroller som är gjorda av revisionsteamet. Och våran avdelning har en specialistfunktion där det jobbar IT-revisorer. Det de fokuserar mest på är behörighets settningen i olika typer av finansiella system. Och ser så att det är rätt satta behörigheter och att de här behörigheterna hanteras på ett korrekt sätt då. Så att man ser då att det är de personerna som får mata in data i programmet som gör det också. Sen så kan de även kontrollera vilket de gör, olika typer av, när det sker filöverföring mellan två system, till exempel med en automatiserad integration mellan… säg ett försystem och ett ekonomisystem. Så ser de att den integrationen är på plats och på så sätt kan uttala sig om att den datan vi har fått från ekonomisystemet den är korrekt och pålitlig och då kan vi börja granska dem. Den korrekta datan.
15	SN	Tackar. Då nästa är Data Currency vilket då handlar om datan är aktuell. Finns det några utmaningar där?
16	R7	Asså där finns det ju också en del utmaningar dock vi jobbar ju alltid med färska data och färsk ska jag säga att ah det är ett år tillbaka i tiden. Och när man är klar med föregående år det stängs ju revisor skriver på. Och då säger ju den att ah den data vi har kollat på den är rätt i den här väsentligheten då. Så jag skulle inte säga att det är något problem som vi brottas med mycket. Den är ju aktuell...som sagt det är ju årsfärsk data så ah… När det gäller våra interna system o så finns det vissa analysverktyg som med utav hjälp av att vi hämtar hem publik data eh och då är det liksom inte vi som står för att datan är korrekt på samma sätt kanske. Så att ah jag skulle att det liksom är ingen frågeställning som verkligen vi granskar eller tänker på så extremt mycket. Utan liksom datat är eller transaktionen är liksom tidsstämplade och är de tidsstämplade inom den perioden som vi ska analysera så ska de vara med där. Annars så ska de inte vara med där.
17	SN	Mm. Ehm ja, nästa är då Data Deduplication och handlar då om det finns dubbletter som refererar till samma entitet. Är det här någon utmaning och hur hanterar ni den i såna fall?
18	R7	Ja. Det är ju en liten utmaning om man jobbar med extremt stora dataset. Och det finns kolumner som beskriver samma typ av data på flera set, eller två olika kolumner beskriver samma data på samma sätt. Då så brukar vi om det skapar problem, be om exporter där det här inte förekommer eller alltså att man bara tar bort en utav kolumnerna. Men oftast så har vi så pass kraftfulla verktyg att det här inte är nått jättestort problem skulle jag säga för just oss.

19	FC	Du säger att ni har kraftfulla system, menar du då att det går igenom någon form av datarengöring först då innan ni får datan och ska börja arbeta med den?
20	R7	Nej, det jag menade med det var att det spelar nästan ingen roll i dem typer utav data hur stora de är, för att våra program klarar av väldigt stora data och ah det bidrar inte till något större problem. Men det beror ju just på att vi arbetar med sådan kort tidsperiod som vi ofta gör och de datasetet som deras program kan få ut inte är så pass detaljerade att det här blir något jätteproblem.
21	SN	Och då är det den sista dimensionen som är Data Consistency, det vill säga i vilken utsträckning som data presenteras i samma format och är kompatibel med tidigare data. Är det något som ni ser utmaningar med?
22	R7	Där är någonting som vi brottas med och är svårt att liksom för oss att påverka eftersom vi hämtar in andras data, som andra bolag äger. Och vi kan ju bara ge rekommendationer hur vi vill att ett dataset ska se ut för att analyserna ska gå så smidigt som möjligt. I vissa fall är det ganska lätt beroende på vilken typ av kompetens den personen man frågar besitter så kan de liksom stuva om datasetet att vi slipper göra det, direkt från systemen. Men när det gäller just bokföring så finns det standardiserade format för hur en bokföring enligt svensk standard, jag tror det till och med bara är Sverige kanske Norden, men det är att man skickar CF-filer. Och det är ju en organisation som står för det formatet CF-gruppen. Så det är ju ett standardiserat format, men när det gäller andra typer utav dataset vi undersöker är det ju väldigt stor skillnad på olika systems exporter och vilken information som följer med från systemen. Så jag skulle säga att det är ett stort problem när man ska automatisera olika typer av processer, men det går ju näst intill alltid att hantera på något sätt, men att det mera är tidskrävande dåra. Så det är väl en aspekt att det är väldigt tidskrävande med olika typer av data från olika typer av system.
23	SN	Så hur hanterar ni det då om ni ändå får det i annat format än det ni vill ha från kunden?
24	R7	Det vi gör då är att vi tvättar datat eh och lägger upp datasetet på det sättet som vi vill och på det sättet som de flesta systemen lägger upp data ehm… så kan man väl säga egentligen att vi transformerar datat till det formatet som vi vill ha det på. Eh och då utför vi kontroller så att vi ser att de datat vi fått in är exakt samma som vi tvättat och stuvat om, men i ett annat format helt enkelt.
25	SN	Amen grymt. Då lämnar vi dimensionerna. Jag tänkte höra med dig om du har hört om det här begreppet ETL som då handlar om Extract Transform Load, och det handlar om vid dataintegration. Så ah jag tänkte egentligen bara höra med dig om ni använder er utav ETL?
26	R7	Ehm nej det är inget begrepp som vi använder oss av direkt.
27	FC	Det handlar just om då man till exempel flyttar data, hämtar ut data från ett system och sen då flyttar över det till er exempelvis och där så sker det ju någon form utav ETL-process för att utvinna datan från kunden för att sedan transformera det och ladda in det hos er. Så det är väl den processen som vi egentligen är lite nyfikna på. Men det kanske inte är något som du personligen är involverad i?
28	R7	Nej, det skulle jag inte säga. Eh, nej det är inget begrepp som vi använder oss utav. Jag funderar på om det kan vara något annat… det är ju oftast någonting som är ute hos just Extract-delen, vi kan ju hjälpa till att extrahera data eller exportera, men det är oftast kunden som gör det. Men transform och load, det är ju liksom våran arbetsuppgift att tvätta datat och göra korrigeringar så att vi kan ladda in dem i det systemet som vi har för analysen.
29	FC	Ja!
30	EU	Yes!
31	SN	Ah men super! Vill du ta över?
32	EU	Ja, ehm utöver de här dimensionerna då och utmaningarna inom det så har vi liksom också identifierat andra utmaningar med datakvalitet. Så jag tänkte även ställa lite frågor kring dem och hur ni hanterar dem, om det är en utmaning för er också.
33	R7	Absolut!
34	EU	Och den första handlar då om urval av data samples. Ehhm och det vi har hittat då i litteraturen är

		att när företag använder sig av stora mängder data så väljer man ofta ut samples som man tränar datan på eller bygger modeller på. Och en utmaning då är att det här samplet inte blir tillräckligt representativt för hela datasetet eller hela populationen. Så använder ni er av samples och är det en utmaning för er?
35	R7	Ja alltså vi har ju en del initiativ tagna för att just ehh göra och bygga olika typer av modeller, för att liksom kunna prediktera olika typer av utfall. Och det är liksom… Det finns inte riktigt den, anser jag då, den kunskapen för ehh vad ett korrekt sample är för att bygga den här datan, eller bygga den här modellen. Men jag skulle väl säga… Alltså generellt sett så tror jag det är ett ganska stort problem, just för att… Ehh man kan ju ta den externa revisionen som ett exempel att där så använder man sig av stickprov när man granskar stora mängder data och just gör samples och då har vi ju olika metoder för hur det samplet görs och det är ju liksom regelrätt om man säger så. Det håller för att kunna säga, ehh i domstol att vi har kollat på hela datat i form av de här stickproven. Sen så tror jag inte att det riktigt följs i den utsträckning som man kanske vill för att uppnå den typen av träffsäkerhet man kanske eftersöker. Men där måste jag ändå också… liksom tidsperspektivet finnas med också att det beror på hur stora datamängder är så kan man liksom inte stickprovsgranska en för stor mängd heller. Och då finns det metoder som plockar ut dem mest väsentliga stickproven liksom och det har vi liksom system för att göra. Men sen gäller det ju att man följer dem till punkt och pricka för att det ska liksom representera hela hela datamängden. Men absolut det skulle jag se som ett… ganska liksom stort problem om man skulle undersöka det lite, absolut.
36	FC	Hur görs dem här stickproven då när datan väljs ut till stickproven? Är det helt slumpmässigt eller har ni någon logik där bakom också som väljer ut datan åt er?
37	R7	Ja alltså det är… vi jobbar egentligen med två stycken, alltså två typer utav, urval. Den ena är liksom helt random sampling där vi randomiserar ut ett stickprov. Men sen så har vi också en, en, en metod som heter Monitor unit sample som viktar dem stora transaktionerna kan man väl säga… ehh dem stora ah transaktionerna kan man väl kalla det och ger dem en större sannolikhet för att bli valda liksom i urvalet då.
38	FC	Ahh okej.
39	EU	Yes, då går vi vidare till nästa utmaning då. Ehh och den handlar om felmarginaler och det är då att i litteraturen så står det att det, att man, alltså att datakvaliteten aldrig kommer bli helt perfekt utan att man får liksom räkna med fel och felmarginaler. Och hur ställer du dig till det att det kommer förekomma felmarginaler i den datan som ni använder? Och hur hanterar ni det då?
40	R7	Ja alltså det, det hanterat ju med hjälp utav ett väsentlighetstal kan man säga när man jobbar med revision. Och det väsentlighetstalet beräknas fram, jag har inte full koll på hur det görs, men sen så alla typer av fel och differenser jämför man mot det här väsentlighetstalet. Ehh och så länge det här väsentlighetstalet är korrekt framräknat med beroende på en rad olika parametrar så kan vi med en viss väsentlighet kunna säga att den här är inte, den här differensen, är inte väsentlig för ehh för det här kontot eller så här. Men ahh vad ska jag säga mer när det gäller just träffsäkerhet och sådär… I vår bransch så är det liksom löst med ett väsentlighetstal och är det så att man går över det här väsentlighetstalet då måste antingen något korrigeras eller att kunden då måste förklara för oss varför den här differensen är på plats. Så dem delarna måste man ju lösa innan man lämnar ifrån sig revisonberättelsen, det slutgiltiga. Men det är någonting som man liksom… vi brottas med skulle jag säga, just i träffsäkerhet och väsentlighet… absolut.
41	EU	Så då är det liksom att det är okej att det liksom blir, att det är här små felen och att det är en felmarginal? Och sen är det ändå en utmaning då att hantera det liksom så att…
42	R7	Ja, att räkna på om den felmarginalen är tillräckligt liten för att man inte ska ta upp den eller inte. Det är väl där utmaningen finns, ehh men absolut så godtar man ju ett fel i datan det gör vi.
43	EU	Aaa, men braa.
44	FC	Bara som en tilläggsfråga där, nu är det väl lite annorlunda för er som sitter med revision just men det är inte så att ni kan välja och alltså ta bort den datan då om den är avviker för mycket från det här talet då som ni räknat fram? Är det okej att bara plocka bort det eller måste ni alltid korrigera det eller?
45	R7	Ja alltså är det över det där väsentlighetstalet måste det korrigeras eller förklaras så att vi ser att differensen inte är så stor som vi först har räknat fram den till.

46	FC	Okej
47	R7	Men att liksom ta bort datan i form av att liksom, vi jobbar ju liksom väldigt lite med att liksom ehh… Ahh vad ska man säga, liksom outline detection och ta bort dem från råmaterialet för att det är ju just dem här outliersen som oftast är väldigt intressanta att undersöka vidare och dem ska vara en del av den mängden man undersöker. Det är snarare så att man söker efter sådana här outliers då och sen så granskar man dem ännu hårdare och dem får en större del av granskningen absolut. Sen kanske det är en lite större skillnad mot hur andra hanterar den typen av observationer. Då förstår den deras modeller så pass mycket att man exkluderar den ur datat, men här så så granskar vi den ytterligare och verkligen presenterar den.
48	FC	Och ser det snarare som något intressant då liksom?
49	R7	Ahh precis och sen när det gäller dem här prediktiva modellerna ehh där jobbar vi idag, i den jag varit involverad i, egentligen bara med olika treaholds. Alltså är ett data, ehh är en datapunkt större eller mindre än det här värdet ah men då exkluderar vi den. Det finns liksom ingen större beräkning varför mer än att det är ett helt osannolikt värde, amen då ska den observationen tas bort då.
50	FC	Okej!
51	R7	Och där finns det kanske lite delade meningar, kring om det är ett korrekt sätt att hantera det på men det är så vi har valt att hantera det i alla fall.
52	EU	Mmm, yes! Nästa utmaning handlar då om datarengöringen och representationen av datan. Och datarengöring kan ju då användas för att lösa problem med saknade värden, syntaktiska fel, alltså om det är ursprungligt låg kvalitet på datan, eller om det är avsaknad av metadata. Och jag tänker… Ni använder er av datarengöring väl, eller?
53	R7	Ja alltså, det skulle jag väl säga att vi på något sätt gör. Ehh kanske inte i våra externa revisionsuppdrag på samma sätt som i våra liksom när ah jag pratar om de här modellerna då. Då är det såklart att vi, vi tvättar data och sådär. Men, men ah vad ska jag säga…
54	EU	Finns det någon utmaning med den processen?
55	R7	Jo men det gör det ju absolut. Utmaningen och risken med den är ju att man missar observationer och data som är väsentliga för modellen och det ändamålet man vill komma fram till. Men ah det läggs inte extremt stor vikt i att undersöka huruvida den datan som man tar bort, ahh det låter ju väldigt dåligt när man säger det men det är väl lite så verkligheten ser ut att det läggs inte så stor kraft på att utvärdera om den här datan bör vara med eller inte som det kanske borde.
56	EU	Nä men hur gör ni för att säkerställa att, alltså att ni behåller ändå en korrekt representation av den datan efter rengöringen alltså att innebörden, alltså att innebörden inte ändras?
57	R7	Mm, alltså när vi jobbar i just den prediktiva modelleringen så jobbar vi ju med ett träningsdataset och ett testdataset och jämför dem två och ser så att träningsdatan representerar den testdatan på ett korrekt sätt. Så att där kan man väl säga att det är träning, test och validering vi utför.
58	EU	Men är det så då att ni liksom har rengjort träningsdatasetet och sen jämför det med det ursprungliga testdatasetet?
59	R7	Ja alltså, emhh ja testdatat blir också på något sätt tvättat. Men den får ju ehh, den får ju den informationen som man har matat in i träningsdatasetet egentligen. I testdatasetet så kan det ju vara outliers och så får vi gå inom observationen, men just tvätta bort så här metadato och sådär det är liksom… Det blir juu… Ahh det jämförs med träningsdatat, så det blir en korrekt liksom representation till… så det representerar testdatan. Så kan väl säga.
60	EU	Mmm
61	FC	Gör ni någonting just när det gäller så här saknade värden eller syntaktiska fel? Har ni något specifikt där som ni gör då, alltså vid datarengöring?
62	R7	Ehh nej egentligen inte. Så som vi har hanterat det i dem modellerna är ju egentligen att vi har… ahh vi tar bort dem helt enkelt. Men också då om informationen inte är fullständig så blir den exkluderad ur träningsdatasetet, det kan man säga.

63	FC	Ja!
64	EU	Yes och hur hanterar ni kostnaderna som uppkommer med datarengöringen? Är det något ni sparar in på eller någonting ni lägger, alltså lägger, väljer att investera i?
65	R7	Nä det skulle jag väl inte egentligen säga. När det gäller allt arbete vi gör ut mot våra externa kunder så tar vi ju liksom betalt för det och just datarengöring kan ju på det sättet om vi får ett väldigt dåligt dataset och det krävs mycket liksom transformering i det så kan man ju också motivera en högre kostnad ut mot kunden. Så i vissa delar så kan det även liksom vara bra att få ett liksom väldigt rådata som måste transformeras och att kunden uppskattar den rengöringen vi gör utav datat och på så sätt vill betala för den tjänsten. I andra fall så vill ju kunderna göra den delen själva och på så sätt liksom spara in hos dem på konsulttimmar. Men det är liksom internet så… ahh nej det kan jag inte riktigt svara på, på samma sätt kanske.
66	EU	Nej det kanske är olika beroende på vilken kund det är. Men ehh okej, vi går vidare till nästa utmaning då och det handlar om ledningens stöd och engagemang. Där vi då har hittat en utmaning med att få ledningens stöd och engagemang angående datakvaliteten då. Men känner du att ni får ett bra stöd från ledningen i processerna med att ha bra datakvalitet?
67	R7	Ahh absolut, det tycker jag och våran liksom bransch är ganska tvingade till att använda sig av dataanalytiker för att liksom ta nästa steg och bli konkurrenskraftiga på marknaden. Så att det är, jag skulle säga… ja alltså självklart vill man ju alltid ha mer stöd för att kunna bygga upp fler liksom interna processer för att kunna effektivisera vårat arbete men samtidigt så tycker jag ändå att vi får absolut det förtroende och stöd som vi behöver för att liksom kunna utföra arbetet på korrekt sätt och sådär, så absolut.
68	EU	Mmm
69	FC	Jag tänker i och med att ni också då jobbar mycket som, alltså som konsulter utåt kunder så ville vi kanske även vinkla frågan dels då till hur det är att få kundernas stöd, ledningens stöd hos kunderna. Till exempel som du nämnde tidigare om det är så att dem ger er data av väldigt dålig datakvalitet, bara ren rå data och att ni då måste bearbeta denna, är det svårt då att få kundens stöd där då att ni behöver dem timmarna till att göra det? Och att de behöver lägga investeringen på det?
70	R7	Ja absolut, alltså det är väldigt mycket sådär fall till fall. Beror lite på liksom mognad och kompetens hos kunderna kring att förstå hur omfattande vissa typer utav arbete är. Men ah generellt sätt så skulle jag säga att det är… många kunder är liksom intresserade av deras IT-miljö och deras liksom deras data och det har väl liksom ökat under åren också. Att många bolag går ju verkligen från det gamla sättet till en mer datadriven verksamhet. Så att ah det är väl ibland man kan känna att vissa kunder inte förstår omfattningen av en del arbete, medan ah andra gånger så kan man få ett jättefint data och de förstår inte hur man kan göra olika typer av analyser för de tycker att det är så komplext för att de besitter en okunskap inom det här ämnet. Men generellt sätt ah ibland kanske, inte alltid.
71	FC	Nä, men där kan jag förstå att det är jätteindividuellt från fall till fall.
72	EU	Ahh
73	R7	Så är det! Och egentligen det har liksom inte någon någon, det spelar nästan ingen roll vilken storlek det är på kunden. Det finns sådana här skräckexempel med jättestora bolag som jobbar väldigt gammeldags kan man säga.
74	EU	Yes, det var alla dem utmaningarna vi har lyckats identifiera så nu tänkte jag höra om det, om du tycket att det finns några fler utmaningar med datakvalitet som vi inte redan har nämnt?
75	R7	Nä men jag tycker att just dem här engelska begreppen täcker väldigt stora delar av vad vi ser för, för, för utmaningar. Nä så att jag tycker att det var jättebra, absolut.
76	FC	Ahh men tack!
77	SN	Super!
78	EU	Ja men vad bra! Om du kommer på något så är det ju bara, så kan du ju bara höra av dig.

Referenser

Alvehus, J. (2013). Skriva uppsats med kvalitativ metod: en handbok, Stockholm: Liber

Anandakumar H., Arulmurugan R., & Suriya M. (2019). Computing and Communication Systems in Urban Development, Cham: Springer

Baesens, B., Bapna, R., Marsden, J., Vanthienen, J. & Zhao, J.L., (2016). Transformational issues of Big Data and Analytics in networked business, *MIS Quarterly*, vol. 40, no. 4, pp. 807-818, Available online: *https://dl.acm.org/doi/10.25300/MISQ/2016/40%3A4.03* [Accessed 26 March 2020]

Batini, C. & Scannapieca, M. (2006). Data Quality Concepts: Methodologies and Techniques, Berlin: Springer

Beebe, N. & Walz, D. (2005). An Empirical Investigation of the Impact of Data Quality and its Antecedents on Data Warehousing AMCIS 2005 Proceedings, Paper 28, Available online: https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1527&context=amcis2005 [Accessed 27 March 2020]

Blake, R. & Mangiameli, P. (2011). The Effects and Interactions of Data Quality and Problem Complexity on Classification, *ACM Journal of Data & Information Quality,* vol. 2, no. 2, pp. 1-28, Available online: https://dl.acm.org/doi/10.1145/1891879.1891881 [Accessed 27 March 2020]

Bloland, P., & MacNeil, A. (2019). Defining & assessing the quality, usability, and utilization of immunization data, *BMC Public Health*, vol. 19, no. 1, pp. 1-8, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 26 March 2020]

Bryman, A. (2016). Social research methods, 5th edn, Oxford: Oxford University Press

Chown, B. (2018). Applying a single source of truth approach to the information needed for Functional Safety 2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC), Available online: https://ieeexplore.ieee.org/document/8569674 [Accessed 29 April 2020]

Churcher, C. (2008). Beginning SQL Queries: From Novice to Professional, New York: Apress

Clarke, R., & Taylor, K. (2018). Towards Responsible Data Analytics: A Process Approach, BLED 2018 Proceedings, Paper 36, Available online: https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1005&context=bled2018 [Accessed 7 April 2020]

Demir, N., (n.d.). Ensemble Methods: Elegant Techniques to Produce Improved Machine Learning Results. Available online: https://www.toptal.com/machine-learning/ensemble-methods-machine-learning [Accessed 18 May 2020]

Dicuonzo, G., Galeone, G., Zappimbulso, E., & Dell'Atti, V. (2019). Risk Management 4.0: The Role of Big Data Analytics in the Bank Sector, *International Journal of Economics and Financial Issues*, vol. 9, no. 6, pp. 40-47, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 8 April 2020]

Dougherty, C. (2016). Introduction to econometrics, 5th edn, Oxford: Oxford University Press

Dupor, S., & Jovanović, V. (2014). An approach to conceptual modelling of ETL processes 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1485-1490, Available online: https://ieeexplore-ieee-org.ludwig.lub.lu.se/document/6859801?arnumber=6859801 [Accessed 15 April 2020]

El Akkaoui, Z., Zimányi, E., Mazón, J. & Trujillo, J. (2011). A model-driven framework for ETL process development Proceedings of the ACM 14th International Workshop on Data Warehousing and OLAP, no. 8, pp. 45-52, Available online: https://dl.acm.org/doi/abs/10.1145/2064676.2064685 [Accessed 3 May 2020]

Elgendy, N. & Elragal, A. (2014). Big Data Analytics: A Literature Review Paper, in P. Perner (eds), *Advances in Data Mining: Applications and Theoretical Aspects*, Cham: Springer International Publishing, pp. 214-227

Emmanuel, I., & Stanier, C. (2016). Defining Big Data Proceedings of the International Conference on Big Data and Advanced Wireless Technologies, no. 5, pp. 1-6, Available online: https://dl.acm.org/doi/abs/10.1145/3010089.3010090 [Accessed 10 May 2020]

Fan, W. (2015). Data Quality: From Theory to Practice, *ACM SIGMOD Record*, vol. 44, no. 3, pp. 7-18, Available online: https://dl.acm.org/doi/abs/10.1145/2854006.2854008 [Accessed 15 April 2020]

Fernandes, N. A., & Wagh, R. (2019). Quality Assurance in Big Data Analytics: An IoT Perspective, *Telfor Journal*, vol. 11, no. 2, pp. 114-118, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 15 April 2020]

Fisher, D., DeLine, R., Czerwinski, M. & Drucker, S. (2012). Interactions with Big Data Analytics, *Interactions*, vol. 19, no. 3, pp. 50-59, Available online: https://dl-acm-org.ludwig.lub.lu.se/doi/abs/10.1145/2168931.2168943 [Accessed 27 April 2020]

Fass, N. (2018). CFOs are Making Data and Analytics Top Priorities, Strategic Finance, 1 October, p. 9. Available from: LUSEM Library website http://www.lusem.lu.se/library [Accessed 26 April 2020]

Gaikwad, S., Nale, P., & Bachate, R. (2016). Survey on Big data Analytics for digital world 2016 IEEE International Conference on Advances in Electronics, Communication and Computer Technology (ICAECCT), pp. 180-186, Available online: https://ieeexplore.ieee.org/document/7942579?arnumber=7942579 [Accessed 7 May 2020]

Google. (n.d.). Google Charts, Available online: https://developers.google.com/chart/interactive/docs/gallery/histogram [Accessed 15 May 2020]

Gupta, A. K., Singhal, S., & Garg, R. R. (2018). Challenges and Issues in Data Analytics 2018 8th International Conference on Communication Systems and Network Technologies (CSNT), Communication Systems and Network Technologies (CSNT), pp. 144-150, Available online: https://ieeexplore.ieee.org/abstract/document/8820251 [Accessed 15 April 2020]

Heinrich, B., Hristova, D., Klier, M., Schiller, A. & Szubartowicz, M. (2018). Requirements for Data Quality Metrics, *Data and Information Quality*, vol. 9, no. 12, pp. 1-32, Available online: https://dl.acm.org/doi/abs/10.1145/3148238 [Accessed 10 April 2020]

Heinrich, B., & Klier, M. (2011). Assessing Data Currency a Probabilistic Approach, *Journal of Information Science*, vol. 37, no. 1, pp. 86-100, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 10 April 2020]

IBM. (n.d.). IBM Knowledge Center: Types of Constraints, Available online: https://www.ibm.com/support/knowledgecenter/SSEPGG_11.1.0/com.ibm.db2.luw.admin.dbobj.doc/doc/c0020149.html [Accessed 15 May 2020]

Ijab, M. T., Surin, E. S. M., & Nayan, N. M. (2019). Conceptualizing Big Data Quality Framework from a Systematic Literature Review Perspective, *Malaysian Journal of Computer Science*, pp. 25-37, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 10 April 2020]

Informatica. (n.d.). What is Data Analytics?, Available online: https://www.informatica.com/services-and-training/glossary-of-terms/data-analytics-definition.html [Accessed 7 May 2020]

Jacobsen, D. I. (2002). Vad, hur och varför: om metodval i företagsekonomi och andra samhällsvetenskapliga ämnen, Lund: Studentlitteratur

Jesiļevska, S. (2017). Data Quality Dimensions to Ensure Optimal Data Quality, *Romanian Economic Journal*, vol. 20, no. 63, pp. 89-103, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 16 April 2020]

Kokemueller, J. (2011). An empirical investigation of factors influencing data quality improvement success AMCIS 2011 Proceedings - All submissions 154, Available online: http://aisel.aisnet.org/amcis2011_submissions/154 [Accessed 17 April 2020]

Lucas, J., Ishfaq, R. & Raja, U. (2014). How Clean is Clean Enough? Determining the Most Effective Use of Resources in the Data Cleansing Process 35 International Conference on Information Systems, Available online: https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1047&context=icis2014 [Accessed 16 April 2020]

Marjani, M., Gani, A., Hashem, I. A. T., Siddiqa, A., Yaqoob, I., Nasaruddin, F., & Karim, A. (2017). Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges, *IEEE Access*, vol. 5, pp. 5247–5261, Available online: https://ieeexplore.ieee.org/document/7888916 [Accessed 20 April 2020]

Morbey, G. (2013). Data Quality for Decision Makers, Erkrath: Springer Gabler

Oates, B. J. (2005). Researching information systems and computing, London: Sage Publications Inc

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data Quality Assessment, *Communications of the ACM*, vol. 45, no. 4, pp. 211–218, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 5 April 2020]

Raheem, N. (2019). Big Data: A Tutorial-Based Approach, New York: Chapman and Hall/CRC

Runkler, T. (2012). Data Analytics: Models and Algorithms for Intelligent Data Analysis, München: Springer Vieweg

Ryen, A. (2004). Kvalitativ intervju: från vetenskapsteori till fältstudier, Malmö: Liber ekonomi

SAS. (n.d.). Big Data Analytics, Available online: https://www.sas.com/en_us/insights/analytics/big-data-analytics.html [Accessed 15 May 2020]

Shatnawi, Q. M., Yassein, M. B., Abuein, Q., & Nsuir, L. (2019). Big data analytics tools and applications: survey In Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems (DATA '19) Association for Computing Machinery, Available online: https://dl.acm.org/doi/abs/10.1145/3368691.3368741 [Accessed 5 May 2020]

Sidi, F., Shariat Panahy, P. H., Affendey, L. S., Jabar, M. A., Ibrahim, H., & Mustapha, A. (2012). Data quality: A survey of data quality dimensions 2012 International Conference on Information Retrieval & Knowledge Management, Information Retrieval & Knowledge Management (CAMP), pp. 300–304, Available online: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6204995 [Accessed 16 April 2020]

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods, *Journal of Business Research*, vol. 70, pp. 263–286, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 25 April 2020]

SmartBear. (2020). RowCount Property, Available online: https://support.smartbear.com/testcomplete/docs/reference/project-objects/items/stores/table/rowcount.html [Accessed 15 May 2020]

Souibgui, M., Atigui, F., Zammali, S., Cherfi, S., & Yahia, S. (2019). Data Quality in ETL process: A preliminary study, *Procedia Computer Science*, vol. 159, pp. 676-687, Available online: https://www-sciencedirect-com.ludwig.lub.lu.se/science/article/pii/S1877050919314097?via%3Dihub [Accessed 1 May 2020]

Tam A., & Kwan I. (2018). Data Quality in Asset Management - Creating and Maintaining a Foundation for Data Analytics, in J. Mathew, C. Lim, L. Ma, D. Sands, M. Cholette, & P. Borghesani (eds), *Asset Intelligence through Integration and Interoperability and Contemporary Vibration Engineering Technologies*, Cham: Springer, pp. 567-574

Taskin, N., Pauleen, D., Intezari, A. & Scahill, S. (2019). Why are leaders trusting their gut instinct over analytics? And What to do About It, *NZ Business + Management*, vol. 33, no. 3, pp. 10-11, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 20 April 2020]

Watson, H. J. (2014). Tutorial: Big Data Analytics: Concepts, Technologies, and Applications, *Communications of the Association for Information Systems*, vol. 34 , no. 65, pp. 1247-1268, Available online: https://aisel.aisnet.org/cais/vol34/iss1/65/ [Accessed 14 April 2020]

Yong, K. K., Shafei, M. S., Sian, P. Y., & Chua, M. W. (2019). Review of Big Data Analytics (BDA) Architecture: Trends and Analysis 2019 IEEE Conference on Open Systems (ICOS), pp. 34–39, Available online: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8975710 [Accessed 6 May 2020]

Yu, Wenyuan. (2013). Improving data quality: data consistency, deduplication, currency and accuracy, PhD thesis, Department of Informatics, Edinburgh University, Available online: http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.615389 [Accessed 25 April 2020]

Zhang, R., Sadiq, S., & Indulska, M. (2019). Discovering Data Quality Problems: The Case of Repurposed Data, *Business and Information Systems Engineering*, vol. 61, no. 5, pp. 575–593, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 20 April 2020]