LUNDS UNIVERSITET
Ekonomihögskolan

Institutionen för informatik

Begränsningar kring IT-verktyg inom Big Data management

Kandidatuppsats 15 hp, kurs SYSK16 i Informatik

Författare: Christian Dahlberg
Rikard Funck

Handledare: Umberto Fiaccadori

Rättande lärare: Benjamin Weaver
Osama Mansour

Begränsningar kring IT-verktyg inom Big Data management

ENGELSK TITEL: Limitations surrounding IT tools within Big Data Management

FÖRFATTARE: Christian Dahlberg, Rikard Funck

UTGIVARE: Institutionen för informatik, Ekonomihögskolan, Lunds universitet

EXAMINATOR: Christina Keller, Professor

FRAMLAGD: maj, 2020

DOKUMENTTYP: Kandidatuppsats

ANTAL SIDOR: 100

NYCKELORD: Big Data, management, IT-verktyg, begränsningar

SAMMANFATTNING (MAX. 200 ORD):
En genomgående exponentiell ökning av Big Data sätter idag press och ställer krav på organisationer. Att effektivisera användningen av IT-verktyg inom processer blir alltmer viktigare för att bekämpa och överkomma problem som kan uppstå under förvaltning av stora datamängder i alla olika stadier av datahanteringen. Syftet med den här studien är därför att undersöka vilka begränsningar de som är verksamma inom Big Data upplever kring de IT-verktyg de använder. Med syftet i åtanke genomför vi därför intervjuer med respondenter och organisationer som arbetar med IT-verktyg inom Big Data management. Våra resultat påvisar att det framförallt handlar om organisatoriska faktorer som bakomliggande problem i de begränsningar som presenterats för oss, vilka sammanfattningsvis omfattar kostnad, kunskap, ställningstagande mot förändringsarbete, volymhantering och Data Governance, men även tekniskt funktionella begränsningar som spårbarhet, kompatibilitet och datakvalitet.

Innehållsförteckning

Tabeller

1 Introduktion

1.1 Bakgrund

De flesta företagen, stora som små, använder och behandlar idag större mängder data i deras dagliga arbeten. Framstegen inom datateknologi och den snabba tillväxten av internet har under flera år medfört många förändringar för data. Teknologiska framsteg har blivit smarta, gjort världen till en mindre plats och gett människan mer tillgänglighet med banbrytande innovationer. Vi har gått från att enbart kunna utbyta data genom fysiskt bemötande (*word of mouth)* till att nu med tillgång till internet kunna få svar på vilken fråga som helst inom någon minut. Framsteg inom teknologier som sensorer, molnbaserade tjänster och maskininlärning har resulterat i en transformerad industri där samtliga faktorer bidragit till en mycket snabb tillväxt av stora volymer data (Raheem, 2019). Data genereras nu från alla möjliga håll, och för det har vi *Internet of Things* (IoT), sociala medier och diverse system att tacka, även smarta hem, smarta bilar, smarta kylskåp och fler smarta produkter. Dessa teknologier har utvecklats för att gynna samhället och förbättra levnadsstandarden för olika individer och organisationer, men den här mängden data bidrar dock inte enbart med positiva faktorer, utan även utmaningar för företag att tackla. Lakoju och Serrano uttalade, (2017) angående data, frasen "*Världens mest värdefulla resurs är nu data*", som även Yi et al. (2014) klassificerat som *den digitala oljan (Big Data).* Frasen publicerades i *Economist* och argumenterar för varför man bör lägga fokus på data och observera det växande intresset inom industrin och akademin (Lakoju et al., 2017). All denna data måste lagras någonstans för oss att ta del av, och även hanteras på ett sätt som är hållbart för aktören som innehar den. Detta sätter press på kraven för systemen och processer för att hantera mängden data, vilket i sin tur medför en varierande men kraftfull kostnad för respektive företag.

För att tackla problem och utmaningar inom Big Data management kan företag ta hjälp av olika IT-verktyg i sina processer för att effektivisera sina steg inom den generella datahanteringen, vilka alla kan se olika ut beroende på företag och bransch, och minimera det ekonomiska utfallet. Detta gör att valet och användningen av dessa IT-verktyg spelar stor roll i hur effektiv datahanteringen inom företaget är för att kunna utvinna värde ur sin data och därav hålla sig konkurrenskraftiga på marknaden, och att dessa IT-verktyg är korrekt utvecklade för att tackla de problem som organisationerna ställs inför.

På senare tid har Big Data-paradigmet skapat turbulens inom många branscher och har i grunden förändrat hur företag driver och fattar beslut i sin helhet. Organisationer har kunnat identifiera betydelsen av att se vilka potentiella medel det finns inom sin genererade data, och kunna dra slutsatser utifrån datan för att effektivisera sina beslut och produkter. Det är jättebra. Men svårigheter kvarstår för företag att exempelvis praktiskt kunna ta del av den insikten. En undersökning gjord av Russom (2013) visar att svagt affärsstöd (investering) och möjligheten att omfatta nya designparadigmer är några exempel på nyckelproblem som kan uppstå hos företag som vill uppnå en bra struktur inom Big Data-hantering. En väletablerad grund i hantering av Big Data, gällande processer ner på IT-verktygs-nivå, är därför att föredra för att minimera risken för att misslyckas (Russom, 2013).

Future Generation Computer Systems rapporterar i en journal av Dobre och Xhafa (2014) att världen producerar omkring 2,5 kvintiljoner byte data varje dag; där 90% av denna data lagras i ostrukturerat format. Detta påvisar varför det är viktigt att kunna ta hjälp av IT-verktyg för att effektivt hantera dessa datamängder; att snabbt kunna strukturera och kategorisera sin data, undgå redundans, och generellt sett minimera kostnader och att eventuella problem uppstår (Dobre et al., 2014). Oavsett varifrån Big Data genereras och mellan vilka parter datan delas, kvarstår utmaningen att hantera den på ett optimalt sätt som ger värde för organisationen, vilket är mer aktuellt idag än någonsin när datamängderna når så massiva nivåer. Lämplig användning av IT-verktyg för databehandling och -hantering kan utvinna data som bidrar med ny kunskap som i sin tur underlättar för företag att jobba effektivt, ligga i framkant och ständigt proaktivt kunna reagera på nya möjligheter och utmaningar i tid (Chen et al., 2013).

1.2 Problemområde

I och med den höga subjektiviteten i arbetssätt bland organisationer, och även fast området Big Data varit i rampljuset i flertalet år, så måste organisationer ständigt arbeta med hur man effektivt hanterar sina datamängder på ett optimalt sätt (Sivarajah et al., 2017). Detta innefattar bland annat hur processer utspelar sig i varje stadium av datahanteringen, vilka IT-verktyg och instrument som används inom dessa, hur väl dessa används för att överkomma problematik och hur väl de implementeras i ett företags arbetsflöde (Sivarajah et al., 2017). Ungefär hälften av all data som struktureras används vid beslutsfattande, och procentandelen av den ostrukturerade datan som analyseras lägger sig under en (1) procent (DalleMule et al., 2017). Detta visar att processer kan förbättras; att spendera mindre resurser och precisera denna investering för att uppnå högre nivåer av användbara data.

Viktigt att nämna är också den största och mest omfattade utmaningen: tillräckliga resurser och kunskapen om var resurserna är bäst lämpade för att ge störst positiva effekt (Sivarajah et al., 2017). För att tackla dessa nämnda komplexiteter krävs mer än enkla statistiska analyser (Zhang et al., 2015). Fler effektiva, mer skalbara och flexibla tekniker och IT-verktyg behövs, och att dessa används på ett genomtänkt sätt för att kunna hantera dessa betydande datamängder ur ett långsiktigt perspektiv (Zhang et al., 2015).

Luckor i kunskapen eller dålig erfarenhet kring datahantering medför problem och utmaningar som är svåra att ta itu med i efterhand. Som samhälle utvecklar vi ständigt mängder med IT-verktyg med syfte att hantera problem; men vet vi egentligen om, och i sådana fall till vilken grad, dessa IT-verktyg löser problemen som kan uppstå inom Big Data management? Även om dagens strategier, teknologier och IT-verktyg används för att tackla problemen så kommer mängden data som idag anses ’stora’ inom en snar framtid vara ’små’ om utvecklingen kring Big Data ökar i samma utsträckning som tidigare, vilket ställer vidare krav på utvecklingen och användningen av dessa teknologier (Sivarajah et al., 2017). Magnituden av problemen kring datahantering ökar på grund av detta eftersom kostnader och resurskrav blir mer påtagliga då alla procedurer blir mer krävande (Sivarajah et al., 2017). Detta motiverar vikten av att ha välfungerande och skalbara IT-verktyg inom datahanteringsprocesser, och att dessa löser den problematik som kan uppstå inom organisationerna. Gantz och Reinsel hävdade (2012) att det i år (2020) ska ha genererats över 40 zettabyte (40 biljoner gigabyte) data. Det är alltså mycket som talar för hur oerhört viktiga dessa problem är att ta itu med inom respektive organisation, som på grundläggande nivå kräver att IT-verktyg med verksamma funktionaliteter utvecklas och används korrekt inom organisationen. Detta som hjälp inom etablerade processer för företag att adoptera och använda sig av.

Tidigare forskning på området tar upp svårigheter och problem inom den generella hanteringen av Big Data; hantera stora volymer data, kunna lagra den, kunna utvinna värde ur den, och göra allt detta på ett kostnadseffektivt sätt (Adiba et al., 2016; Almeida, 2017; Barnaghi et al., 2013; Chen et al., 2013; Sivarajah et al., 2017; Vaghela, 2018). Adiba et al. (2016) tar upp jämförelser mellan olika typer av databaser och Big Data management plattformar, och poängterar flertalet problem som kan uppstå i varje steg av Big Data-livscykeln. Resterande studier (Almeida, 2017; Barnaghi et al., 2013; Chen et al., 2013; Sivarajah et al., 2017; Vaghela, 2018) tar upp problem på ett generellt och övergripande plan, men fåtalet talar om hur väl organisationer tar hjälp av befintliga IT-verktyg för att tackla dessa problem. Hur ser egentligen användningen av dessa ut i praktiken? Den här undersökningen kommer att djupdyka inom ämnet och utforska vilka IT-verktyg inom processer för Big Data management som finns hos organisationer; om och hur de används, och till vilken grad de löser identifierade problem, eller om det i annat fall finns begränsningar kring användningen av dessa IT-verktyg som omöjliggör eller försvårar detta. Vi vill sammanfattningsvis alltså fokusera på vilka begränsningar företag som hanterar Big Data finner inom de IT-verktygen de använder, och dess bakomliggande faktorer. Problemställningen är därför aktuell för majoriteten av de stora företag som använder sig av Big Data och, genom studerande av den här undersökningen, potentiellt hålla sig mer konkurrenskraftiga och förhoppningsvis öka sitt värde på marknaden.

1.3 Forskningsfråga

Utifrån ovanstående aspekter kring bakgrund och problemområde formulerar vi forskningsfråga med följdfråga som följer:

Vilka begränsningar hos IT-verktygen för hantering av Big Data upplever arbetande inom dessa processer? Vad är anledningen till att dessa begränsningar kvarstår?

1.4 Syfte

Syftet med den här undersökningen är att undersöka vilka begränsningar aktörer verksamma inom hantering av Big Data upplever kring användandet av IT-verktyg, och varför i sådana fall dessa begränsningar kvarstår inom respektive organisation. Detta görs i hopp om att fylla ett kunskapsgap kring där förbättringar går att utföra, och även i sin tur inspektera vad de bakomliggande faktorerna är för att skapa mer medvetenhet kring dessa. Även varför begränsningarna runt användningen av dessa IT- verktygen är aktuella inom organisationerna, för att i längden ta steg närmare optimala processer inom hantering av Big Data.

1.5 Avgränsningar

Trots att studien inkluderar ett fåtal exempel på IT-verktyg som frekvent används inom Big Data management är huvudsyftet inte att identifiera nya så kallade IT-verktyg. Studien kommer även exkludera begränsningar som baseras på GDPR eller andra juridiska faktorer. Fokus kommer alltså ligga på begränsningar kring användningen av IT-verktyg som de verksamma inom företag använder i deras processer för Big Data-hantering.

2 Litteraturgenomgång

I det här avsnittet definierar vi termen Big Data som kommer ligga som grund för hela studien, och även dess olika dimensioner som kommer vara utgångspunkterna för de IT-verktyg vi vill analysera. Därefter definierar vi även Data Management och vad vi anser 'IT-verktyg' vara i vår kontext. Sen presenterar vi tidigare forskning för problem och utmaningar som kan uppstå inom alla olika stadier av Big Data-hantering, innan vi exemplifierar ett antal IT-verktyg skapade för att hjälpa organisationer och motverka potentiella utmaningar. Till sist fastställs en sammanfattning som innehåller en fördjupad problemformulering där tidigare problemområde ligger som grund.

2.1 Begränsningar

2.1.1 Definition

Vi utgår från den officiella definitionen av en "begränsning", där en begränsning definieras som en restriktiv svaghet; något med brist på kapacitet, alltså en oförmåga att utföra något som den aktuella, utan denna begränsning, skulle ha kapacitet till (Dictionary, n.d.). I den här studien kommer vi även anse att begränsning är en direkt respons på att ett IT-verktyg inte löser ett visst problem.

2.2 Big Data

2.2.1 Definition

För att förstå hantering kring Big Data är det viktigt att förstå begreppet Big Data och vad som skiljer det från vanliga data. Big Data är först och främst en term som beskriver stora volymer data som kan kategoriseras i tre större strukturer; strukturerade data (exempelvis relationsdatabaser), semistrukturerade data (nyckel-värdetal som i Mongo databaser) och ostrukturerade data (exempelvis textfiler, filmer, bilder och email) (Raheem, 2019). Den här mängden data är givetvis enorm, men det är inte volymen av datan som är viktig - utan vad man gör med den (Raheem, 2019). Big Data definieras inte enbart av stora mängder data, utan även av ytterligare dimensioner, som förklaras djupgående under nästa rubrik, som bör beaktas där skillnader mellan vanliga data och Big Data påvisas (Davenport, 2012).

En annan aspekt är att se Big Data som en synonym för *Data Analytics,* likheterna är tydliga men det finns andra aspekter som kan definiera Big Data. Tidigare har Big Data definierats av tre dimensioner som fastställts i en modell vid namn *The 3 Vs of Big Data* som innehåller *volume*, *variety* och *velocity* (McAfee, 2012). Den här modellen har senare utvecklats till den mest väldefinierade modellen, *The 5 Vs of Big Data*, som även inkluderar *value* och *veracity*, vilka fastställer de mer genomgående egenskaperna av Big Data (McAfee, 2012). Det är inom dessa fem punkter som skillnaden mellan vanliga data och Big Data hittas.

2.2.2 Dimensioner av Big Data

Volume, eller volym, är som tidigare nämnt, i stora drag det som generellt sett har kännetecknat Big Data (Raheem, 2019). Den här dimensionen hänvisar till de stora mängder data som genereras varje dag, vare sig det är från exempelvis *Internet of Things* (IoT), kommunikationsdata eller transaktionsdata (Raheem, 2019). Antalet data som skapats varje dag har sedan 2012 legat på 2.5 exabyte (ungefär 1 000 000 gigabyte), och det antalet sägs dubblas var 40:e månad (McAfee, 2012). I de flesta organisationerna skiljer sig även data från data, där vissa data samlas in för allmän datalagring (exempelvis personuppgifter) medan annan specifikt för analys (exempelvis köpbeteenden) (McAfee, 2012). Alla olika sorters data gör att omfånget av Big Data påverkar kvantifieringen, vilket leder oss in till nästa dimension.

Variety, eller variation, avser alltså de olika formaten och filtyperna av den Big Data man samlar in, och hur oföränderlig datastrukturen är (Adiba et al., 2016; Russom, 2011). Big Data handlar alltså inte enbart om volymen. Tidigare har vi sett outnyttjad strukturerade data samlas in och bara förvarats utan att något värde utvunnits från den, vilket vi nu ser ändring på genom att företag väljer att analysera sin data istället för bara förvara den (Russom, 2011). Detta är exempelvis data från leveranskedjeapplikationer genom användning av RFID, textdata från callcenter-applikationer, geospatial data och GPS-avkänning inom logistik och även användningsdata från medarbetare inom processer på deras företag (Russom, 2011).

Man kan numera betrakta oss alla som levande datageneratorer. Företag måste ta hand om semi- och ostrukturerade data (*XML, RSS feed, human language text)* från källor som är nya för dem, vilket genast ställer krav på mjukvara och teknologi att kunna hantera detta (McAfee, 2012). Den minskade kostnaden av elektronisk utrustning bidrar till ett större antal användare som genast ökar utvinningen av data genom dessa produkter (McAfee, 2012). Detta har hjälpt oss in i en ny era: en där stora mängder av digital information existerar om praktiskt taget varje ämne för företag att ta del av och skapa värde från (McAfee, 2012).

Velocity, eller hastighet, syftar på med vilken frekvens företag samlar in data på, och hur snabbt man bearbetar datan för att göra den tillgänglig för användning (Adiba et al., 2016; McAfee, 2012). Detta kan ofta vara viktigare än själva volymen data, eftersom företag kan göra snabbare beslut om datan samlas in mer frekvent. Realtidsinsamling av data tillåter företag att agera mer aggressivt och före sina konkurrenter, vilket bidrar till en högre konkurrenskraft (McAfee, 2012). Det sägs att 90% av världens data genererats under de senaste två åren, vilket visar att den ökade hastigheten av dataströmmar har bidragit till utvecklingen av de mängder data som vi ser idag (SINTEF, 2013).

Value, eller värde, har under senare tid identifierats som ännu en dimension av Big Data. Datan som genereras har i sin ursprungliga form vanligtvis ett relativt lågt värde gentemot dess volym eftersom inget värde har utvunnits ännu (Oracle, n.d.-a). Datan är alltså inte användbar förrän det värdet upptäcks, och behöver därför genomgå analys och hantering innan ett värde genereras (Oracle, n.d.-a). Att lagra strukturerade-, semistrukturerade- och ostrukturerade datamängder är en hyfsat enkel uppgift för företagen, men den stora utmaningen är att extrahera tillförlitliga data och utvinna affärsvärde av den (Raheem, 2019).

Veracity**,** eller sanningsenlighet, omfattar ens tillit till den datan man samlat in eller otillförlitligheten till de datakällorna man använt (Gantz et al., 2012). När det gäller osäkerhet kring numeriska data kan företag i flesta fall kombinera flera mindre tillförlitliga källor eller ren matematik som tillsammans bygger upp en tillit och en mer exakt datapunkt (Rubin et al.,

2013). Men när det gäller osäkerhet kring textdata så ökar komplexiteten markant, eftersom denna data oftast härstammar från sociala medier och i synnerhet från personer (Rubin et al., 2013). Därför är den här datan både osäker i uttryck och innehåll, och måste därför analyseras på ett annat sätt än numeriska data.

På senare år har det tillkommit fler dimensioner som hjälpt definiera Big Data, vilka vi valt att utelämna i den här undersökningen då vi anser att dessa dimensioner inte ännu är fullt etablerade på samma skala som de vi ovan definierat. Dessa ytterligare dimensioner omfattar exempelvis *Visualization* och *Variability* (Ferguson, 2012). *Visualization*, eller visualisering, behandlar hur man konceptuellt presenterar Big Data i ett format som människor förstår, och *Variability* är helt enkelt bara en annan form av *Variety* (Ferguson, 2012).

2.3 Data management

2.3.1 Definition

Data management, eller datahantering, är ett brett ämne som innefattar insamling och lagring, men även behandling och leverans av data (Russom, 2013). Discipliner inom datahantering omfattar exempelvis datalagring, -integration, -kvalitet, -säkerhet med mera. Dessa är procedurer som har med information att göra; vad och hur man ska beröra information som man samlat in eller själv genererat.

Big Data management, eller hantering av Big Data, omfattar liknande discipliner och procedurer som ovanstående *Data management* men applicerat till det som definierar Big Data (Russom, 2013). Big Data kan, som ovanstående definierat, omfatta annorlunda struktur, innehåll och typ än vanliga data, vilket ställer krav på discipliner inom hantering av just denna sorts data och att kunna bemöta den mångfalden. Mjukvarulösningar inom BDM tenderar att innehålla verktyg och lösningar som är mer mottagliga för fler sorters data, större mängder, och som kan bearbeta denna data snabbare (Russom, 2013).

Data behöver först och främst samlas in på något vis, och detta är första steget i hanteringen av Big Data (Almeida, 2017). Den här datan lagras då i organisationens databaser (oftast via molntjänst när det handlar om stora volymer), innan den går vidare för att bearbetas, beräknas och sedan analyseras i hopp om att kunna identifiera värdefull information ur datan (Almeida, 2017). När data har samlats in, blivit lagrad, bearbetad och analyserad, och till slut omdefinierad till värdefulla data så visualiserar företag ofta sina resultat för andra att ta del av, innan de utnyttjar sina insikter (Almeida, 2017). De här olika faserna förklaras i nästa avsnitt, där även ett par IT-verktyg som kan finnas i de olika faserna exemplifieras.

2.4 Problem och utmaningar inom Big Data management

Användning och hantering av Big Data för med sig ett antal problem och utmaningar för organisationer. Dessa problem kan inom processer av Big Data-hantering skapa begränsningar för organisationer om de inte tas itu med. De utmaningar som vi har valt att ta upp täcker generell datahantering och visar olika sorters problem som kan uppstå i olika faser av Big Data management. I introduktionen av den här studien presenterade vi att förändringsarbete, exempelvis anpassning till nya designparadigmer, kan vara ett exempel på problem som kan uppstå hos företag som vill gå från en suboptimal Big Data-hantering till en bättre (Russom, 2013). Det är ett organisatoriskt problem som kan vara svårt att övervinna i och med att det handlar om flertalet medarbetare, där alla tillsammans måste jobba i symbios och i samma riktning. Här är det även viktigt att poängtera att för att utveckla och implementera infrastruktur för hantering och bearbetning av Big Data så krävs skicklig personal med kunskaper om Big Data (Sivarajah et al., 2017).

I det här avsnittet tar vi upp tre huvudutmaningar; (1) utmaningar relaterade till dimensioner eller egenskaper hos själva datan; (2) utmaningar som uppstår vid bearbetning och analys av data som består av att fånga upp, tolka, och slutligen presentera sin data och sitt slutresultat; (3) utmaningar relaterade till hantering av Big Data, exempelvis när du får åtkomst till, hanterar och styr din data (Sivarajah et al., 2017).

2.4.1 Datarelaterade problem och utmaningar

Den massiva ökningen av storskaliga volymer data har, som vi tidigare konstaterat, blivit alltmer aktuell på senare år (Adiba et al., 2016; Almeida, 2017; Barnaghi et al., 2013; Vaghela, 2018). Detta bidrar med nya utmaningar för redan etablerade IT-verktyg att hantera, och kan komma att kräva justering eller att nya metoder för att hantera problemen utvecklas (Almeida, 2017; Sivarajah et al., 2017; Vaghela, 2018). Likväl kan olika former och kvalitéer på data indikera att heterogenitet känns naturligt inom Big Data, vilket påvisar vikten och utmaningen av att ha system och IT-verktyg som kan förstå och hantera den sortens data (Almeida, 2017; Chen et al., 2013; Labrinidis et al., 2012). Det är även viktigt att betona den frekvens som data samlas in på, görs tillgänglig och tiden för bearbetning (Adiba et al., 2016). Alltså hur snabbt en organisation kan bearbeta icke-homogen data så den blir tillgänglig för användning, antingen genom skapande av nya data eller uppdaterande av befintliga data Chen et al., 2013). Detta ställer krav på IT-verktygen att hantera och kan lätt orsaka långsiktig problematik om det inte sköts på ett korrekt sätt (Adiba et al., 2016; Chen et al., 2013). Innan man analyserar sin data krävs det att man har tillförlitlighet för den (Gantz et al., 2012). Här uppstår problem med att mäta noggrannheten och datans potentiella användningsgrad för analys, i och med bias, tvivel, rörigheter eller dylikt kring den (Sivarajah et al., 2017). När man pratar om höga hastigheter av data som strömmar in för organisationer, så kräver det även att organisationer hittar metoder för att klara problemen med att lagra och hantera, samt slutligen utnyttja data på ett kostnadseffektivt sätt (Vaghela, 2018; Sivarajah et al., 2017).

2.4.2 Processrelaterade problem och utmaningar

Att extrahera heterogena data från olika källor, och lagra denna för värdeskapande är en stor komplexitet som kan bidra med problem som aldrig tidigare förekommit inom datainsamling och lagring (Abawajy, 2015; Sivarajah et al., 2017). Om IT-verktyg eller applikationer inom organisationer inte kan hantera detta begränsar det hastigheten som information kan

extraheras och lagras på, och vidare med vilken frekvens som värde kan uppnås (Abawajy, 2015; Chen et al., 2013). Efter extrahering och lagring av datan, behöver man korrekt extrahera och färdigställda data från en insamlad pool av storskaliga, ostrukturerade data (Sivarajah et al., 2017). Detta förespråkar flera inom Big Data som en möjlighet att öka effektiviteten på (Chen et al., 2013). Detta, genom att utveckla nya procedurer, för att snabbt och enkelt kunna utvinna rengjorda data från större pooler, och öka den generella effektiviteten av Big Data management (Chen et al., 2013). När ens data är extraherad och rengjord kan organisationer uppleva problem kring integrationen av sin data, vilket behandlar aggregering och kategorisering av datan, som är högst subjektiv beroende på bransch och vilken data som används (Chen et al., 2013; Labrinidis et al., 2012; Sivarajah et al., 2017). Den här sortens data kan, beroende på typ, givetvis sakna naturligt bindande information för aggregering, vilket kan bidra till svårigheter och problematik för organisationer i det här stadiet av datahanteringen (Sivarajah et al., 2017). När organisationer i slutändan vill analysera sin data, och erhålla värde från den, behöver de främst iaktta skillnaden mellan att hantera vanliga data och Big Data (Sivarajah et al., 2017). Man måste också poängtera att äldre metoder för dataanalys främst brukade handla om att lösa komplexiteten av sambandet mellan data inom relationsdatabaser (Sivarajah et al., 2017). I och med ökande krav på lagringsutrymme och effektivitet så krävs det nyare, utvecklade metoder och IT-verktyg för att hantera Big Data, för att ge organisationer möjlighet att få ett maximalt monetärt affärsvärde från sin data. Samtidigt behöver organisationer förutspå, genom analys, vad som kan komma att ske i den kort- eller långsiktiga framtiden. Detta innebär en utmaning för organisationer, för att hålla sig konkurrenskraftiga (Chen et al., 2013).

2.4.3 Managementrelaterade problem och utmaningar

I introduktionen nämnde vi hur vår nuvarande digitala tidsålder bidrar till en massiv ökad mängd data som idag genereras överallt, hela tiden (Raheem, 2019). Detta föranleder stora problem gällande sekretess, och hur vi behandlar människors integritet blir en stor utmaning för samtliga organisationer. Detta är extra tydligt för data som överförs via nätverk av Big Data-applikationer, som exempelvis sensorer för platsbaserad information (GPS) (Yi et al., 2014). Vi nämnde tidigare hur smarta stadsmiljöer med ett antal sensoriska enheter idag samlar in mycket större mängder med data om aktiviteter än tidigare, vilket även försvårar situationen och utgör ett ännu större integritetsproblem för medborgare (Barnaghi et al., 2013). Gällande integritet så måste även säkerhet av data involveras. Problem kring säkerhet av Big Data skiljer sig inte mycket från vanliga data, utan handlar helt enkelt om enorma mängder data som måste skyddas från vanliga skadliga program, som länge varit ett ständigt växande hot mot den generella datasäkerheten (Yi et al., 2014). Sofistikerad infrastruktur kring datasäkerhet kan vara en stor utmaning för organisationer då mängden datatyper, volymen av datan, och kravet på snabba processer växer (Demchenko et al., 2013; Almeida, 2017; Vaghela, 2018). I och med den komplexa naturen hos Big Data, generell brist på struktur och beblandade filtyper, så kommer en viktig utmaning för organisationer vara att kategorisera och mappa sin data sedan de extraherats och lagrats (Almeida, 2017; Sivarajah et al., 2017). Detta är ett viktigt steg för att säkerställa kvalitén på utfallet från användningen av sin data, något som i annat fall kan stå organisationen dyrt (Almeida, 2017; Sivarajah et al., 2017). Kostnaderna kommer alltså alltid vara en stor utmaning inom Big Data management (Vaghela, 2018). Framförallt i och med den ständigt ökande mängden heterogena data som idag är aktuell, vilket bidrar till en ökad efterfrågan på beräkningsresurser inom organisationer för att hantera detta (Vaghela, 2018). Forskare hävdar att utveckling av en strategi för att avsevärt minska kostnaderna är akut för att effektivisera hanteringen av dessa mängder komplexa data (Sivarajah et al., 2017). Operativa utgifter, kunskap och kostnader kring

databehandling kan för organisationer sätta käppar i hjulet för hur de implementera teknologiska lösningar och IT-verktyg kring detta (Al Nuaimi et al., 2015; Vaghela, 2018). I sig är detta ett problem och en långsiktig utmaning.

2.5 Big Data management

Som bakgrund till det praktiska problemet och de generella utmaningar som finns kring Big Data så krävs det etablerade IT-verktyg för att lösa problemen och driva processer inom hantering framåt kring Big Data. Dessa IT-verktyg går att hitta i samtliga olika faser inom Big Data, vare sig de är utvecklade internt och är skräddarsydda inom organisationen, eller standardiserade verktyg hämtade från större, mer etablerade operatörer.

Efter att ha gått igenom ett antal journaler och topplistor från IT-hemsidor har vi exempel på populära IT-verktyg som kontinuerligt används som hjälpmedel inom Big Data management. En journal om Big Data skriven av Fernando Almeida (2017) visar en tankekarta över ett antal IT-verktyg som används inom Big Data, detta och ett antal topplistor på populära verktyg har legat som grund för de verktyg vi valt att ta upp (DataFlair, 2019).

Vi har strukturerat upp faser av Big Data management i följande logiska ordning (se 2.5.1 - 2.5.6), och därefter exemplifierat ett IT-verktyg per fas som kan användas inom denna för att motverka problem och utmaningar som kan förekomma. Dessa är också delvis sorterade efter popularitet (Edupristine, 2017). Strukturen håller en logisk ordning där kategorierna följer verkligheten; från hur data extraheras, transformeras och laddas upp emot en databas (*ETL*) i rätt sorts format och struktur, till hur data lagras för att sedan bearbetas, beräknas och därefter analyseras. I slutändan visualiseras resultatet från den analyserade och beräknade datan, med en avslutande beskrivning av vad en plattform innebär ur ett tekniskt perspektiv. Verktygen definieras kortfattat med användningsområde och på vilket sätt de används.

2.5.1 Extrahering

Att kunna hantera data kräver att man innehar data, vilket innebär att organisationer behöver extrahera den här datan på något sätt, såvida de inte genererar den själva. Extrahering av data är processen att samla in data från olika källor för att sedan lagra det på ett eller flera ställen så att vidare åtgärder kan utföras (Sivarajah et al., 2017). Den extraherade datan kommer alltså i ett senare skede ge värde åt organisationen, vilket betonar vikten av att rätt data samlas in just för detta ändamål eftersom den insamlade datan oftast är ostrukturerad i den här fasen (Sivarajah et al., 2017). Nedan exemplifierar vi ett IT-verktyg som används inom Extrahering, som tillsammans med *Transform* och *Load* bildar ETL (*Extract, Transform, Load*), och kan ses som en av de första faserna av datahantering (Sivarajah et al., 2017):

Talend är en open-source dataintegrerings-plattform som erbjuder mängder av olika produkter och services inom integration, datakvalité, molntjänster och framförallt Big Data (Chand, 2019). Detta verktyg kan förenkla och hjälpa till med integreringen av den data som samlas in inom organisationen, och även att åstadkomma värde av den insamlade datan (Chand, 2019). Man pratar här mycket om ETL som är den process man använder sig av vid integrering av data. Den består av tre viktiga steg, vilket omfattar att data först extraheras och samlas i en singulär lagringsplats för att sedan transformeras, valideras och reduceras i hopp

om att ta bort redundant data, och slutligen läsas in i någon form av lagringsutrymme (Talend, n.d.-a).

2.5.2 Lagring

En databas är i sin renaste form ett system som lagrar data på ett strukturerat sätt och gör datan lättillgänglig för användaren (Cambridge Dictionary, n.d.). När man lagrar data använder man databaser för att i ett senare skede kunna använda denna data främst för att generera värde för organisationen (Sivarajah et al., 2017). Almeida förklarar i en artikel (2017) att databas- och lagringsverktyg inom Big Data har ett dubbelt syfte; en infrastruktur för att köra analysverktyg inom men samtidigt en plats för att lagra och hämta Big Data från. Några exempel att tänka på vid val av lagringsverktyg är typ och storlek av sin data, potentiella tillväxtförväntningar och vilken sorts databas som organisationen använder (SQL/NoSQL) (Almeida, 2017). Nedan har vi exemplifierat ett populärt databassystem som används inom Big Data management:

MongoDB är i grunden en distribuerad NoSQL-databas, men den är dokumentbaserad och byggd för molntjänster och moderna applikationer (MongoDB, n.d.-a). *MongoDB* använder sin egen JSON-baserade syntax med nyckel-värde par, som körs i symbios med majoriteten av de stora programmeringsspråken för att kommunicera med databasen (MongoDB, n.d.-b).

2.5.3 Processing

Processverktyg är grundläggande för Big Data-hantering och omfattar ofta en snabb motor för Big Data-analys som integrerar olika bearbetningstekniker, exempelvis maskininlärning och grafbehandling (Almeida, 2017). Att processa data innebär alltså att, genom användning av dessa processverktyg och -tekniker, ta data utan kontext, så kallad rå data, och transformera den till meningsfull information (Talend, n.d.-b). Informationen kan sedan användas av organisationer för att genomföra analyser på och därefter kan ta viktiga beslut från (Talend, n.d.-b). Nedan har vi exemplifierat ett populärt processverktyg som främst används inom Big Data management, och förklarat några av dess grundfunktioner:

Hadoop MapReduce är ett ramverk inom *Apache Hadoop* som gör det lättare att bygga applikationer för att beräkna stora mängder data. Själva modellen går ut på att en viss input-data delas upp i olika självständiga bitar som vidare används som input i en så kallad ‘*reduce*’-process (Apache Hadoop, n.d.-b). Reduceringsprocessen kan exempelvis bestå av ett genomförande av olika summeringar, exempelvis en summering på hur många som har samma namn i ett kluster. Syftet är att reducera en bit så att endast meningsfulla data återstår (Apache Hadoop, n.d.-b).

2.5.4 Analys

När datan är färdigprocessad är den redo för nästa steg: analys. Målet med att analysera ens data förklaras enligt Fan et al. (2014) i två olika steg; att från ens data utveckla effektiva metoder för att kunna förutspå framtida observationer när man vid nästa tillfälle extraherar data, och för det andra få insikt i subkluster av ens data. Kort sagt så behövs analys av datan för att hitta mönster och strukturer dolda i subkluster av data som du som organisation har extraherat i ett tidigare skede (Fan et al., 2014). När du är klar med att analysera din

strukturerade, bearbetade data kan du dra stor nytta av att identifiera gemensamma egenskaper hos datan och sådant som kan användas och gynna ditt företag ekonomiskt. Nedan har vi exemplifierat ett IT-verktyg som hanterar just analys av en färdigprocessad data:

Oracle Data Mining (ODM) är en komponent inom *Oracle Advanced Analytics* och den erbjuder kraftfulla analytiska algoritmer som hjälper till med att förutse smarta investeringar eller förutspå framtida kundbeteenden från sin data (Oracle, n.d.-b). ODM erbjuder även ett grafiskt *drag and drop* gränssnitt för analytiker att ta del av. Genom det här verktyget kan analytiker analysera sin data och visualisera resultatet för att i god tid kunna reagera på förändringar på marknaden och hålla sig konkurrenskraftiga (Oracle, n.d.-b).

2.5.5 Visualisering

Att visualisera data och/eller information innebär att presentera den i ett läsbart format (Sivarajah et al., 2017). För att göra det mer begripligt använder man sig ofta av olika grafiska verktyg såsom grafer och diagram av diverse slag (Sivarajah et al., 2017). Detta är en mycket viktig process eftersom man utifrån denna presentation ska kunna ta viktiga beslut inom organisationen. Problem kan här lätt uppstå på grund av de många dimensionerna som Big Data besitter (Sivarajah et al., 2017). Nedan har vi exemplifierat *Tableau*; beskrivit mjukvaran och den del inom *Tableau* som hanterar visualisering:

Tableau är ett program som kombinerar olika mjukvaror för att leda data från förberedelse till visualisering (Tableau, n.d.). Den officiella dokumentationen för Tableau (n.d.) beskriver mjukvaran som en plattform bestående av ett antal olika subverktyg; *Tableau Prep, Tableau Desktop, Tableau Online* och *Tableau Server*: *Tableau Prep* förbereder data för analys genom användandet av ett gränssnitt som förenklar arbetet markant. *Tableau Desktop* erbjuder också ett *drag and drop*-gränssnitt där man kombinerar data i olika format för att erhålla olika analysresultat. *Tableau Online* är en molnbaserad tjänst med ett grafiskt gränssnitt som möjliggör delning mellan kollegor och kunder av de analyser som utförs. Slutligen *Tableau Server* som har samma funktion som *Tableau Online* med skillnaden att det körs på en server och inte via molnet, och är tänkt för interna analyser (Tableau, n.d.).

2.5.6 Plattformar

En plattform beskrivs av Bottcher (2018) som en miljö, ofta ett operativsystem, bestående av flertalet olika teknologier där applikationer kan köras och användas. Ett bra exempel på en stor, molnbaserad plattform är Apache Hadoop. Apache Hadoop kan kort beskrivas som en mjukvarumiljö i kombination med hårdvara som i sin tur möjliggör att andra relaterade applikationer installeras och körs inom den miljön (Apache Hadoop, n.d.-a). Plattformen förenklar även processering och analys av data (Apache Hadoop, n.d.-a). Andra bra exempel på populära plattformar är *Microsoft Azure, Amazon Web Service*s och *Google Cloud* (Computerworld, 2020; Datamation, 2020).

För att kunna bearbeta Big Data, med hjälp av dessa plattformar och ytterligare verktyg, behöver man oftast hjälp av olika programmeringsspråk och mindre supportmjukvaror. Detta kan vara allt från olika API:er, universellt använt format som JSON, databaslagring inom SQL och framförallt NoSQL (vilka är mycket populära inom Big Data-hantering på grund av sin enkelhet och skalbarhet) (Almeida, 2017). Även machine-to-machine möjligheten som ger Big Data-miljöer möjlighet att kommunicera mellan enheter (Almeida, 2017). Detta ger

företag de stöd de behöver och erbjuder stora möjligheter inom Big Data-hantering (Almeida, 2017).

2.6 Sammanfattning av litteraturgenomgången

Syftet med den här undersökningen är att ta reda på vilka begränsningar de som är verksamma inom Big Data management upplever kring de IT-verktyg de använder inom sina processer. *Big Data* är en term som ofta beskriver stora volymer data, men definieras av Raheem (2019) enligt tre olika kategorier; strukturerad-, semistrukturerad- och ostrukturerade data. Den här datan kräver hantering innan den kan användas och förhoppningsvis ge värde för organisationen. Hanteringen av Big Data, Big Data management, definierar Russom (2013) som discipliner och procedurer för att samla in, lagra, bearbeta och leverera olika typer av Big Data. I och med definitionen av Big Data så kräver BDM discipliner inom Big Data-hanterings alla faser (*Extrahering, Lagring, Processing, Analysering* och *Visualisering*) som kan bemöta den mångfald av filtyper, struktur och innehåll som Big Data kan innefatta (Russom, 2013).

2.6.1 Sammanfattning av problem och utmaningar inom Big Data-hantering

I tabellen nedan sammanfattar vi olika identifierade problem och utmaningar som kan uppstå inom faserna av datahantering, vilka alla på något sätt kan behöva tas itu med, som potentiella begränsningar kan sätta stopp för. Problemen är sorterade utifrån tre (3) typer som tidigare kategoriserats som (1) Datarelaterade-, (2) Processrelaterade-, och (3) Managementrelaterade problem och utmaningar.

Tabell 1: Sammanställning av identifierade problem och utmaningar

Typ	Problem och utmaningar	Litteratur
Data	Hantera stora volymer data	*(Adiba et al., 2016; Almeida, 2017; Barnaghi et al., 2013; Sivarajah et al., 2017; Vaghela, 2018)*
Data	Hantera hög variation på data	*(Almeida, 2017; Chen et al., 2013; Labrinidis et al., 2012; Vaghela, 2018)*
Data	Bibehålla hög hastighet och frekvens av insamling och förberedande av sin data	*(Adiba et al., 2016; Chen et al., 2013; Vaghela, 2018)*
Data	Tillförlitlighet och tillit av sin data	*(Gantz et al., 2012; Sivarajah et al., 2017)*
Data	Värdet av den data man samlar in	*(Abawajy, 2015; Sivarajah et al., 2017)*
Process	Fastställa effektiv extrahering och lagring av sin data	*(Chen et al., 2013; Sivarajah et al., 2017)*
Process	Fastställa processkraft för bearbetning och beräkning av data	*(Chen et al., 2013; Sivarajah et al., 2017)*
Process	Aggregera och integrera sin data på ett effektivt sätt	*(Chen et al., 2013; Labrinidis et al., 2012; Sivarajah et al., 2017)*
Process	Snabba och effektiva analyser av sin data	*(Chen et al., 2013; Sivarajah et al., 2017)*
Management	Bibehålla hög integritet på sin data för alla inblandade	*(Yi et al., 2014; Barnaghi et al., 2013)*
Management	Säkerställa hög säkerhet mot hotfulla externa mjukvaror	*(Almeida, 2017; Demchenko et al., 2013; Vaghela, 2018; Yi et al., 2014;)*
Management	Data Governance	*(Almeida, 2017; Sivarajah et al., 2017)*
Management	Minska kostnader kring alla stadier av sin datahantering	*(Al Nuaimi et al., 2015; Sivarajah et al., 2017)*

3 Metod

I det här avsnittet kommer vi gå in på våra tillvägagångssätt gällande varje steg av vår undersökning. Vi kommer introducera och motivera för läsaren vårt generella metodval och beskriva våra procedurer för datainsamling. Vidare kommer vi motivera urvalet av respondenter för vår empiriska studie, beskriva denna studie och även inkludera faktorer som etik, moral, transkribering och validitet av denne.

3.1 Metodval

Den här undersökningen har utförts genom ett iterativt arbetssätt där varje del blivit ifrågasatt av oss själva innan tillagda i texten. Varje tillagd del ska ha god motivering till varför, och enbart inkluderas ifall det har potential till att bidra till undersökningen. Efter att ha studerat Big Data management som ämne, och gjort förarbete på var vi anser det största kunskapsgapet funnits så behövde vi avgränsa oss till ett område som går att undersöka inom vår tidsomfattning. Då vi identifierat problematiken kring huruvida väl IT-verktyg inom processer för Big Data management ger det stöd hos verksamma inom Big Data de behöver, så började vi gå djupare in i problemen som kan uppstå inom den generella hanteringen av Big Data. För att se hur användningen av IT-verktyg inom Big Data fungerade i praktiken behöver vi undersöka hur de som är verksamma inom Big Data management upplever bruket av sina IT-verktyg, och se vilka begränsningar de finner runtom dessa, för att få en förståelse kring vad som behövs förbättras. Detta för att generellt hjälpa att ta steg i rätt riktning mot en mer effektiv datahantering inom organisationer. För att uppnå detta använde vi en kvalitativ undersökningsmetod i form av intervjuer.

3.1.1 Kvalitativ undersökning

En tillbakablick på undersökningens syfte visar att målet med den här undersökningen är att påvisa vilka begränsningar de som är verksamma inom Big Data upplever genom användning av IT-verktygen, vilket visar att det är kvalitativ information från företagen vi strävar efter. För att samla in ny, kvalitativ data så utförde vi en undersökning i form av intervjuer. Eftersom vi söker data som kan vara unika för respektive företag så lär större delen av information vara ny för oss, vilket Jacobsen (2002) uttrycker ingår i en kvalitativ undersökning; att vara öppen och mottaglig för ny information. Vikten av en kvalitativ undersökning när man undersöker specifika teorier eller områden är något Jacobsen (2002) lägger tyngd på för att skapa en fördjupad förståelse kring området i fråga. Detta handlar i vårt fall om problematik inom Big Data management och de IT-verktyg inom dessa processer.

Det är svårt att göra en uppskattning om vilka begränsningar olika företag upplever genom användningen av sina IT-verktyg utan att göra en undersökning kring ämnet. Respektive intervjuobjekten kan ha olika synpunkter och åsikter kring användandet av IT-verktygen, och kanske uppleva olika begränsningar, om ens några, vilket kräver att vi som intervjupersoner är öppna för flertalet olika tankar och tolkningar. En kvalitativ undersökning hjälper forskare att fånga upp detaljer och sanningsenlig information med hög validitet genom att låta intervjuobjekt framföra tankar, åsikter och tolkningar öppet utan några specifika riktlinjer (Jacobsen, 2002). Det här arbetssättet öppnar dörren för en semistrukturerad intervju vilket

tillåter oss att dynamiskt strukturera våra intervjuer inom vår omfattning och avgränsning; att kunna ändra ordning på frågor, och tillägga och/eller exkludera frågor under intervjuns gång, allt för att kunna hålla en öppen diskussion för att maximera insamlandet av kvalitativ information (Jacobsen, 2002). Detta gör att vi som står bakom intervjun och undersökningen är öppna för alla möjliga svar (Jacobsen, 2002).

3.2 Datainsamling

3.2.1 Litteraturinsamling

Informationen under litteraturgenomgången är hämtad från ett fåtal etablerade journal-plattformar med hög validitet såsom *LUBSearch* och *Google Scholar*, samt från enstaka böcker relevanta för vår frågeställning och metod. Dessa plattformar har gett oss artiklar och journaler främst utifrån sökorden Big Data, utmaningar (*challenges*), datahantering/management, IT-verktyg (*tools*) inom Big Data, och ofta utifrån sökningar som innehöll kombinationer av dessa.

3.2.2 Intervjuguide

Frågorna är skapade och färdigställda med litteraturgenomgången som bakgrund. Detta gör i sin tur vår studie lättare att replikera, ifall mer forskning inom det aktuella ämnet skulle göras i framtiden, eftersom vi behåller en röd tråd och en artad struktur genom hela den litteraturgenomgång som intervjustrukturen och -guiden bygger på. Vi har valt det här sättet för att i ett senare skede kunna ställa vår empiriska fakta emot vår litteraturgenomgång, för att se hur förankrade organisationerna är inom faserna av datahantering såväl som deras IT-verktyg. Även hur väl dessa IT-verktygen ger det stöd för organisationerna som de behöver, eller om de i annat fall visar begränsningar, samt följder och effekter av dessa begränsningar för verksamheten Vi använder litteraturen som en grund att stå på, och för att ge bakgrundsfakta och kunskap om ämnet Big Data management, samt problematik kring ämnet, allt för att undersöka vår frågeställning angående begränsningarna kring dessa IT-verktygen inom empirin.

Som tidigare nämnt under 3.1.1 Kvalitativ undersökning har vi utfört en semistrukturerad intervju, där en intervjuguide etablerats innan första mötet. Vår intervjustruktur, utöver faktumet att den är semistrukturerad, etablerades med syftet att ha en tydligt uppdelad intervju som följer litteraturgenomgångens struktur med öppna frågor och ett dynamiskt upplägg. Med dynamiskt upplägg menar vi att intervjun vid behov finslipats efter respektive intervju ifall någon fråga eller del av intervjun ställde krav på revidering. Detta kände vi var en viktig del av vår empiriska studie då eventuella infallsvinklar eller nyanser kunde identifieras under loppet av respektive intervju.

Tabell 2: Intervjuguide för huvudfasen av intervjun

Del	Undersökning	Exempelfrågor	Nyckelord
Bakgrund och Big Data management	Bakgrund till organisationens datahantering och syn på deras kunskap om processerna inom varje steg	• Hur skulle du definiera "Big Data"? • Förklara en översikt av er allmänna datahantering för Big Data	*Bakgrund, Big Data, management*
IT-verktyg	Identifiering av IT-verktyg för att se ifall de har några specifika IT-verktyg inom deras datahantering för att bygga nästa fas på.	• Berätta för oss om vilka IT-verktyg ni använder inom era processer för Big Data management • Inom vilka specifika steg (eller faser) inom Big Data management använder ni dessa verktyg?	*IT-verktyg, Processer, Faser, Management*
Begränsningar	Se över deras IT-verktyg och dess kända begränsningar, samt effekterna av dessa.	• Hur skulle du definiera "begränsningar"? • Beskriv eventuella begränsningar inom era respektive IT-verktyg som används inom era processer för Big Data management • Hur tror du att ett plötsligt behov av att hantera en ökad mängd data skulle påverka användningen av era respektive IT-verktyg?	*IT-verktyg, Begränsningar, Big Data management*
Framtiden	Avslutningsvis se över framtiden för deras datahantering och följder av begränsningarna.	• Vilka är de kortsiktiga utmaningarna för dig och din Big Data-ledningsverksamhet? • På vilket sätt tror du att begränsningarna av de presenterade IT-verktygen kan påverka er på lång sikt? • Är det något du vill tillägga som du tycker att vi missat att ta upp?	*Framtiden, Utmaningar, Begränsningar*

Tanken med vår semistrukturerade intervju var att undgå en snäv och statisk struktur. Med litteraturgenomgången som grund strukturerade vi upp intervjuguiden med en top-down-struktur och utvecklade den med en logisk, dynamisk struktur. Först identifierade vi uppdelningen av intervjun, för att sedan sekventiellt följa upp under intervjuns gång. Intervjuns struktur var upplagd med samma upplägg som litteraturgenomgången, och involverade frågor kring deras generella datahantering, följt av eventuella IT-verktyg inom deras processer för hantering av Big Data, och slutligen om de anser att de upplever några begränsningar hos dessa. Utifrån detta kunde vi i nästa steg fila på våra frågor; undvika ja och nej frågor, samt frågor med förutfattade svar. Dessa öppna exempelfrågor låter respondenten tala fritt kring ämnet, och håller intervjun öppen som en diskussion snarare än en utfrågning, och låter oss bidra med eventuella följdfrågor (Jacobsen, 2002). Vi har även inkluderat en

kolumn med nyckelord som kortfattat beskriver vad respektive del kommer att handla om (se *Nyckelord).* Slutligen adderade vi en kolumn (se Undersökning) som beskriver för läsaren våra intentioner, vad uppdelningen betyder, och vad de ställda frågorna vill undersöka. Detta gjordes mest för att på ett snabbt och enkelt sätt kunna ge bakgrundsinformation till läsare, intervjuperson och respondent. Det är viktigt att ha i åtanke att intervjuguiden bara används som en *guide*, och behöver inte följas till punkt och pricka, vilket går hand i hand med en semistrukturerad intervjustruktur (Jacobsen, 2002).

3.3 Urval av respondenter

För att besvara vår frågeställning om begränsningar kring IT-verktyg inom Big Data-hantering krävdes det att vi hittade verksamma som arbetar inom organisationer som hanterar Big Data i sin vardag. Majoriteten av organisationerna är stora, väletablerade företag inom Sverige, med en respondent från ett internationellt företag som är väl insatt i Big Data. Samtliga respondenter som intervjuats är föregående eller nuvarande anställda, med arbetsroller inom Big Data management, som till vardags jobbar med IT-verktyg som behandlar enorma mängder heterogena data. De olika respondenterna arbetar även inom olika branscher. Detta har varit ett aktivt beslut i och med att IT-verktyg som används inom Big Data känns oberoende av bransch, vilket gör studien mer intressant om det är ett bredare spektrum vi undersöker.

För att välja respondenter utgick vi främst från ett bekvämlighetsurval vilket innebär att man kontaktar personer som man känner sen tidigare (Jacobsen, 2002). Om kunskapen runt Big Data eller användningsområdena kring de personer vi kontaktade inte gav oss tillräckliga respondenter använde vi en metod som kallas *Snöbollsmetoden,* vilken går hand i hand med ett bekvämlighetsurval, och menar att de man kontaktar rekommenderar andra personer som skulle vara relevanta för våra ändamål (Jacobsen, 2002). Snöbollsmetoden gav oss tre av fem (3/5) respondenter medan bekvämlighetsurvalet två (2). Respondenterna är väletablerade i ämnet och jobbar dagligen med IT-verktyg och processer som hanterar stora datamängder. Samtliga respondenter fick en intervjuguide sig tillsänd några dagar innan. Dessa innehöll strukturen på intervjun vilket uppskattades enormt då de fick tid att förbereda sig.

3.3.1 Respondenter

Nedan följer en tabell över samtliga respondenter, information kring dessa och kring intervjun i sig, och i vilken appendix transkriberingen av hela intervjun finns.

Tabell 3: Information om respondenter och intervjuer

Företag	Namn	Arbetsroll	Intervjutyp	Tid	Datum	Appendix
Företag A	Respondent A	R&D Engineer	Videosamtal	34 min	2020-04-15	A
KPMG	Mats Dahl	Director Digital Data & Analytics	Videosamtal	49 min	2020-04-16	B
SEB	Pia Carlsson	Business Glossary Manager	Videosamtal	30 min	2020-04-21	C
Företag B	Respondent B	Director Consulting Services BI & Analytics	Videosamtal	43 min	2020-04-22	D
Företag C	Respondent C	Application Architect	Röstsamtal	45 min	2020-04-24	E

3.4 Bearbetning av empiri

Respektive intervju erbjöd oss möjligheten att reflektera, och ändra något ifall något var oklart eller otydligt. En dynamisk intervjustruktur lät oss vara öppna med våra frågor och hur vi bemötte respondenternas olika svar, vilket vi ansåg ökade variansen och gav mer intressanta svar. Frågorna kunde alltså komma att korrigeras vartefter och skräddarsys för respektive respondent för att få ut så mycket information som möjligt på varje fråga.

3.4.1 Inspelning

Samtliga intervjuer har hållits på distans med hjälp av mjukvarorna *Microsoft Teams och Discord*. För att spela in intervjuerna har vi använt OBS (*Open Broadcast Software*), som ofta används vid streamingtjänster, för att spela in ljudet på respektive mjukvara. Efter att intervjun utförts och spelats in, la vi tid till att analysera och reflektera över intervjun, och diskuterade eventuella funderingar eller synpunkter som kan ha kommit upp. När vi reflekterat, gjort anteckningar och sammanfattat vad vi kommit fram till så har varje intervju transkriberats på en dator i separata dokument, utifrån de inspelade samtalen, för att sedan läggas in under Appendix i vår rapport. Detta har vi gjort måttfullt och detaljrikt för att säkerställa att samtliga data från intervjun finns med, vilket även hjälper till att hitta nyckelinformation för analys (Jacobsen, 2002).

3.4.2 Transkribering

Tillvägagångssättet för transkriberingen har behållit samma struktur genom hela processen, och ett fåtal frågor har konfigurerats efter respektive intervju då vi utfört en semistrukturerad intervju. Varje intervju transkriberades ordagrant från början till slut, ord för ord, men kan ha

genomgått vissa förändringar gällande meningsuppbyggnad, eller borttagning av redundans, för att öka läsbarheten och flödet av texten. Här är det viktigt att poängtera att inga ändringar gjordes som ändrade andemeningen i respondentens uttalanden, utan utfördes för ökad kvalité. Vi har försökt hålla all transkribering begriplig och lättläst, med en struktur som tydligt visar vilka som är intervjuare och vilka som är respondenter. Tidpunkter har även noterats, i vilken ordning, och hur lång tid samtliga delar av intervjuerna har tagit. Vi anser att resultatet av total transkribering, alltså transkribering som inte reducerats till "det viktigaste" eller till "några typiska bitar", bidrar med en klarare översikt. Helt enkelt för läsaren att få möjlighet att gå igenom hela intervjun.

3.4.3 Analys

Analysen har utförts med hjälp av en mall som vi satt ihop som en modell, vilken behandlar varje del av intervjun. Mallen utgår från en analysteknik framlagd av Flick et al. (2004), och är specifikt anpassad för semistrukturerade intervjuer. Metoden vi utgått ifrån är avsedd att uppmuntra analytiker att utveckla sina egna analysmetoder (Flick et al., 2004). Detta tog vi till oss och inkluderade delar som vi kände passade in för vår typ av intervju, samt exkluderade onödiga processer som helt enkelt var olämpliga. Valet av analysmetod beror mycket på ens mål med studien, på vilka frågor ens intervju består av, och inte minst på hur mycket tid och resurser som finns tillgängliga (Flick et al., 2004).

Tabell 4: Analysmodell (tabell) för empirisk bearbetning

Steg 1	**Materialbaserad formation av analytiska kategorier**
	Läsa materialet (transkriberade semistrukturerade intervjuer)
	Fastställa analytiska kategorier utifrån materialet som grund för hela analysen
Steg 2	**Montering av analytiska kategorier**
	Noggrant deskriptiva rubriker för kategorisering
	Testa med en del av empirin om fastställda kategorier är användbara
Steg 3	**Fastställa analyserade data från intervjuerna med kategorierna som grund**
	Iterera genom resultatet och klassificera data utifrån fastställda kategorier
	Gå över redundans och minimera mängden data till hanterlig mängd
Steg 4	**Fylla kategorierna med analyserat material**
	Fastställa resultat utifrån analyserat material
	Se likheter, likheter, frekvens av begränsning
	Se hur många problem x begränsning givit respektive organisation
Steg 5	**Detaljerade validering av fastställda resultatet**
	Noggrann validering av resultatet gentemot respektive frågor från intervjun

Modellen är strukturerad utifrån mallen i fem (5) olika steg: **Steg 1** omfattar formation av resultat-tabellens kategorier som man fastställer genom att läsa samtliga transkriberingar och helt enkelt tar ut det viktiga ur ett top-down perspektiv. Detta kan kännas tidskrävande, men

det är fundamentalt eftersom man inte vill förbise något väsentligt från respondentens sida, och framförallt inte blanda in bias från vår egen sida (Flick et al., 2004). För vår del fastställde vi begränsningarna som kategorier eftersom det framförallt var det vi ville fokusera på och analysera med vår frågeställning. Under **Steg 2** monteras de rubriker som ligger som grund för ens kommande analys, här bör man försöka minimera likheter mellan rubrikerna, och verkligen fastställa ytterst deskriptiva kategorier så att de inte överlappar, eller att läsaren kan uppleva dem vaga (Flick et al., 2004). Under **Steg 3** utförs själva analysen. Här bearbetar man resultatet iterativt och klassificerar informationen med ovan definierade kategorier i åtanke. Under det här stadiet av analysen är det viktigt att minimera mängden information man använder för att kunna göra jämförelser, om man så vill, på ett lättare sätt och därefter se likheter eller skillnader intervjuer emellan (Flick et al., 2004). När ens information är analyserad och klassificerad så är det bara för oss att presentera det analyserade materialet. Det är i det här stadiet, **Steg 4**, som analytiker förhoppningsvis vill börja fastställa resultat, som för vår del möjligtvis skulle omfatta likheter och skillnader mellan synpunkter, frekvens på respektive begränsning inom olika organisationer, relationer mellan x och y, antal begränsningar organisationer upplever och identifierar. Slutligen, under **Steg 5** itererar analytiker igenom ens färdiga resultat gentemot respektive frågor i intervjun för att se om det ser lämpligt ut och ger läsaren hög tillförlitlighet och validitet på våra resultat. Här kan målet vara att upptäcka nya hypoteser, testa en hypotes på enskilda fall, att skilja mellan termer, komma fram till nya teoretiska överväganden eller att revidera befintliga teoretiska ramverk (Flick et al., 2004). Allt beror på ens upplägg på intervjuerna och med vilken sorts data man arbetar.

3.5 Validitet och reliabilitet

För att bibehålla en hög kvalité på vår studie krävs en hög validitet och reliabilitet på materialet. Detta har säkerställts genom en utvärdering av vårt empiriska resultat. Utvärdering av resultatet utfördes för att fastställa relevans och giltighet av transkriberingen, som ligger till grund för ett resultat av hög kvalité. Efter färdigställd transkribering har dokumenten skickats till respektive respondent för att låta dem korrigera felaktigheter och eventuella missuppfattningar. Vi inväntade mottagande från respondenterna innan vi kunde fastställa en högre intern validering av materialet (Jacobsen, 2002). För att fastställa en extern validering behövde vi fastställa huruvida vårt resultat gick att generalisera gentemot andra studier (Jacobsen, 2002).

Vi nämnde kortfattat i tidigare avsnitt hur en semistrukturerad intervju tillät intervjuare att göra en intervju med mer dynamisk struktur; att kunna ändra frågor allteftersom inför respektive intervju. Detta är något vi har haft i åtanke för att öka den generella tillförlitligheten på studien och undgå *intervjueffekten*, som innebär att framförandet av intervjufrågorna kan påverka respondentens svar, och därmed sänka en studies reliabilitet (Jacobsen, 2002). I vårt dynamiska tillvägagångssätt har vi försökt att fortfarande behålla samma framförande av intervjufrågorna, oavsett typ och struktur av frågan, för att minimera effekten av denna problematik.

3.6 Etik och moral

Vi hanterade alla respondenter på samma sätt, från att vi etablerade kontakt till att vi var klara med intervjun. Vi använde oss av riktlinjer utfärdade av Oates (2005) för att säkerställa att alla kände sig väl behandlade utifrån ett etiskt perspektiv.

Vi började alltid med att skicka ut intervjun, strukturen och frågorna, i förväg för att försäkra oss om att vi skulle uppfattas på ett seriöst sätt och säkerställa respondenten känsla av förberedelse, samt ge personen chans att läsa igenom frågorna vilket är viktigt för utförandet av en intervju (Oates, 2005). Möjligheten att som respondent vara anonym är något som borde värdesättas högt, både av individen och den organisation denne tillhör, detta för att reducera eventuell stress hos respondenten ifall denne delar känslig information om organisationen (Oates, 2005). I och med detta har vi alltid varit noggranna med att i början fråga respondenterna till vilken grad de vill bibehållas anonyma.

I början av intervjun gav vi alltid tydlig information kring vilka vi är och vad vårt syfte med studien och dess material är. Vidare tillfrågades respondenten om det var i sin ordning att spela in intervjun i transkriberingssyfte. Efter att inspelningen avslutats erbjöds respondenten att vid färdig transkribering, ta del av denna för att skapa en enhetlig syn på intervjun och att undvika missförstånd. Dessa steg togs för att skapa transparens och sätta respondenten i en trygg sits (Oates, 2005).

3.7 Metodreflektion

När man, som vi, valt att genomföra en kvalitativ studie istället för kvantitativ måste man enligt Bryman (2016) fokusera på att bygga äkthet och trovärdighet runt den studie man utför. Detta har vi försökt främja på ett antal punkter för att visa att vi är seriösa i vårt arbete.

Angående respondenterna försäkrade vi oss alltid i förväg om att de förstod materialet väl och kunde ge oss något utifrån det. I den här undersökningen sökte vi oss till personer som jobbat eller jobbar med IT-verktyg berörande Big Data management och/eller personer som innehar en position inom sin organisation där individen har översikt över vilka IT-verktyg man använder. Ett exempel på gallring av potentiella respondenter var när vi kom i kontakt med en individ som verkade kunna väldigt mycket om Big Data och maskininlärning, vilket från början lät attraktivt, men som senare visade sig inte vara användbart då individen inte hade särskilt bra översikt över vilka verktyg man använde.

Avslutningsvis vill vi poängtera att en kvalitativ utgångspunkt var rätt val för oss i och med att vi gör en undersökning inom ett aktuellt ämne som växer för varje dag. En kvalitativ metod gav oss mer öppna svar och kommentarer kring ämnet vilket stärker vår förståelse för forskningsfrågans problematik.

4 Resultat

I den här delen kommer resultatet från vår empiriska undersökning att presenteras. Detta kommer omfatta information från alla intervjuer och inkludera hänvisning till Appendix 1–6 där en intervjuguide och en transkribering av samtliga intervjuer finns samlade. Kapitlet kommer att introducera respondenternas definitioner av både Big Data och vad en begränsning inom en mjukvara är, och sedan vara uppdelad, för att underlätta navigering för läsaren, i de olika kategorierna som identifierats under litteraturgenomgången och låg till grund för varje intervju.

4.1 Definitioner

4.1.1 Big Data

Vi har under tidigare avsnitt poängterat att Big Data består av en uppsjö av olika definitioner och betydelser beroende på vem du frågar. Respondenterna i vår studie definierar Big Data, likt vår teoretiska iakttagelse, på en rad olika sätt. Respondent A definierar Big Data som metoder för att behandla enorma datamängder, och sedan genom användningen av denna data kunna förutspå saker på ett bättre, mer skräddarsytt sätt. Mats på KPMG har en definiering ur ett optimeringsperspektiv, dock en mer volymbaserad definition, och menar att Big Data är användandet av en obegränsad mängd data för en obegränsad mängd olika typer av tjänster, med syftet att dra nytta av och förändra affärsmodeller och verksamheter till det bättre. Respondent B ser Big Data ur två olika perspektiv och menar att bredden på data spelar roll; antingen ett brett dataset med lågt antal entiteter men stort antal attribut per entitet, eller ett smalt dataset med få attribut men väldigt högt antal entiteter, och avslutar med att definiera Big Data som "*svårhanterliga mängder data*". Respondent C tar upp "5 V"-modellen och liknande modeller som definierar dimensionerna av Big Data, och är delvis kluven till tillvägagångssättet kring modellerna. Hen nämner att Big Data kan definieras på olika sätt från person till person. Avslutningsvis definierar Pia på SEB Big Data som möjligheten att skala upp och ner vid behov.

4.1.2 Begränsning

Respondent A definierar en begränsning inom en mjukvara som avsaknaden av funktionalitet hos ett verktyg, alltså att personen istället väljer ett annat verktyg för att utföra en viss funktionaliteten. Respondent A fortsätter och menar att avsaknaden av funktionaliteten innebär en begränsning inom mjukvaran. Mats definition pekar på att ifall verktyg är utvecklade med låg kvalité resulterar det i att diverse processer tar längre tid än de borde, vilket är en begränsning hos det verktyget. Respondent B har som definition att ett verktyg med en begränsning inte löser det problemet man vill lösa. Denne exemplifierar att det handlar om att information inte levereras i tid. Avslutningsvis definierar Pia en begränsning som att ett verktyg inte kan utföra det man förväntar sig, och att man då istället tittar på andra verktyg.

4.2 Big Data Management och IT-verktyg

På det stora hela anser majoriteten av våra respondenter att deras processer för datahantering kring Big Data fungerar bra, framförallt stegen för generering och extrahering (ETL). Man nämner att man ibland upplever svårigheter, problem eller brister i användandet av datan i de olika stegen. Majoriteten använder sig av standardiserade lösningar som *Amazon Web Services* eller *Microsoft Azure*, men poängterar även att *Google Cloud* börjar bli mer och mer populärt. Här berättar Respondent B att användning av internt utvecklade lösningar för extrahering och lagring av data "är som att uppfinna hjulet igen".

Respondent A nämner att det för dem ofta handlar om enorma mängder dataströmmar från tredje part, där flertalet datatyper är involverade, som deras applikationer och internt utvecklade IT-verktyg måste ta hand om. Respondent A ger exempel på att en självproducerad kortfilm, filmad under ett par veckor, gav företaget 500 terabyte data att jobba med.

Våra respondenter poängterar att det framförallt är de operativa, organisatoriska aspekterna som ligger till grund för problemen de upplever hos användningen av deras IT-verktyg, vilket exempelvis kan handla om kunskap om ett visst arbete. Pia på SEB nämner att det inom bankväsendet till större del verkar handla om brist på kunskapen, om vilken data som ska tittas på och användas. Detta håller Respondent B med om och nämner även *traceability* och *lineage*, eller spårbarhet, av sin information som två större faktorer. Inom bankbranschen, fortsätter Pia, genererar man oftast och köper in sin data, men man hämtar även in stora mängder referensdata som läggs i en så kallad *Data Lake*. Pia beskriver en *Data Lake* som en stor uppsamlingsplats för all inhämtad data, som man sedan använder för att göra analyser, API:er och rapporter. Inom organisationens *Data Lake* påpekar Respondent C att de använder ett visst antal kontroller av sin data så att de kan kategorisera data på ett sätt som underlättar värdeskapandet av den. Här talar Respondent B om komplexiteten och svårigheten (kring kunskapen) att veta vilken data från IoT-applikationer som är intressant att skapa värde från, och för detta, och alla andra processer inom en organisations datahantering, menar Respondent B att det krävs applikationer och verktyg.

4.2.1 Standardiserade IT-verktyg

“Buy before build” - Pia, SEB

Både Pia på SEB och Mats på KPMG nämner att de inom organisationen implementerat en policy som strävar efter att undvika alltför många internt utvecklade produkter, och att man i stället ska försöka använda standardiserade lösningar som finns att köpa och hämta digitalt.

Mats på KPMG nämner att de använder *Microsoft Azure* som plattform för att hantera Big Data som en följd av diverse krav från kunder. Han nämner specifikt att de arbetar med de flesta mikrotjänster som Azure erbjuder inom ramen för Big Data. Dessa involverar *Azure Bot Service, Microsoft AI, Azure Machine Learning* och även deras *Text Analytics API*, ett NLP-verktyg (*Natural Language Processing*), samt *Power BI* för visualisering. Vidare poängterar han att Microsofts övriga tjänster och IT-verktyg täcker allt man behöver genom alla deras processer för Big Data management. KPMG använder *Microsoft Azure* (i grund och botten) på grund av affärsmodellen, som Mats anser skiljer sig från andra leverantörer. En ytterligare viktig faktor till varför KPMG väljer *Microsoft* påstår Mats är att de tillhandahåller

infrastrukturen i molnet utan att äga kunddatan. Denna fullständighet som *Microsoft Azure* verkar erbjuda presenteras av Respondent B, som nämner att Microsoft Azure har en fördel att de erbjuder allt genom hela processen, och framförallt att de har *Power BI* som slutverktyg för visualisering. Respondent B berättar att de använder standardiserade IT-verktyg och lösningar genom hela deras organisation, men skräddarsydda konfigurationer inom dessa för att anpassa procedurer till organisationen.

Pia berättar att de snarare köper in och använder redan etablerade IT-verktyg och lösningar än att bygga dessa internt, och nämner att de på SEB använder *Informatica* som huvudverktyg för hela processen för hantering av Big Data. Pia beskriver *Informatica* som ett stort och dyrt företag som ofta implementeras av större företag som har ekonomiska resurser att använda det. Respondent C nämner också *Informatica*, främst Informatica Cloud och Informatica On-Prem, som deras standardiserade integrationsplattformar när det gäller att flytta sin data, vilket intervjupersonen påstår fungera bra.

Respondent C tar även upp *Amazon Web Services* (AWS) som deras primära plattform för Big Data. Denne nämner flertalet tjänster inom AWS, exempelvis *Amazon Lambda*, som Respondent C förklarar som kortlivade funktioner som bara kan exekveras inom en kort tidsrymd, och används för validering av data vid olika checkpoints; *AWS Glue*, som de delvis använder för integration, men även för att skanna sina dataset och hitta anomalier eller att identifiera känslig information; *DynamoDB* som databas, och för att hantera metadata med syfte att främja spårbarhet; och slutligen AWS *Simple Storage Service* (S3) vilket används för att lagra rådata i ett specifikt filformat som gör datan snabbare att hämta. Respondent C förklarar att detta bara är exempel på de tjänster inom AWS som de använder inom deras organisation, men att AWS erbjuder alla IT-verktyg som de kan önska och fyller deras krav på att kunna skräddarsy deras konfigureringar.

4.2.2 Internt utvecklade IT-verktyg

Respondent A nämner att deras verksamhet är väldigt gammal och att de därför förlitar sig på flertalet legacy system, och genom alla 30 åren har de alltid utvecklat interna IT-verktyg, applikationer och lösningar som används till alla processer längs hela deras pipeline. IT-verktygen används främst för att samla statistik över hur mycket lagring och liknande som används vid en given tid, och om hur mycket deras olika avdelningar för datahantering används. Respondent A berättar att det finns väldigt mycket data på statistik, men att de främsta IT-verktygen är mindre script som används till att överföra data från A till B, oavsett datatyp. Hen fortsätter och menar att de, som tidigare nämnts, oftare genererar sin egen data än att samla in referensdata eller data från tredje part. Respondent A avslutar med att nämna att de inte använder några standardiserade lösningar eller verktyg som Hadoop, och jämför sig med större sociala medieföretag eller banker som tar in större mängder persondata. Man påpekar också att man hanterar Big Data, men inte i samma volymer som de större företagen.

4.3 Tekniska begränsningar

4.3.1 Kompatibilitet

En majoritet av respondenterna tar upp kompatibilitet hos deras IT-verktyg som en solklar begränsning, men på olika sätt. Pia nämner att de haft ett 40-tal olika kundsystem som de nu försöker reducera ner till ett. Det handlar ofta om *legacy system* när det är äldre företag med i bilden, vilket gäller för samtliga organisationer vi pratat med. Respondent A nämner att *legacy system* ger större svårigheter att använda verktyg som är kompatibla med skräddarsydda script och internt utvecklade verktyg, och påpekar att de inte är helt dynamiskt kompatibla med andra, nyare verktyg. Respondent A fortsätter med att poängtera att det inte är lätt att uppdatera ett system för att göra det bättre, eftersom risken för att försämra kompatibiliteten med andra verktyg då ökar markant. Mats på KPMG nämner också att deras verktyg och lösningar måste bli mer kompatibla med varandra oberoende på hur datan ser ut och vilken kund det handlar om, allt för att minimera begränsningarna och optimera deras generella datahantering. Här talas det om en fungerande "*common data model*" som ska kunna generalisera användandet av de IT-verktyg KPMG använder och kunna tillämpas i alla olika delar av verksamheten.

4.3.2 Spårbarhet

Två av våra anonyma respondenter, Respondent B och C, upplever spårbarhet som en tydlig begränsning inom deras befintliga verktyg och tjänster. Respondent B nämner att det finns en tydlig brist på hur man kan spåra ens data genom olika processer för datahantering, både i molntjänster och on-premises. De använder sig för närvarande av Microsofts produkter och anser att det ändå är relativt lätt att följa sin data och information. Dock påtalar man problematik inom *Oracle* databaser och *DB2*, samt att de som använder verktyget *Informatica* för sin datahantering kommer att uppleva begränsningar vid spårbarhet inom det verktyget. Samtidigt tar Respondent C upp att det deras organisation främst saknar hos de nuvarande verktygen är just spårbarhet, och poängterar vikten av att förstå vad som hänt med datan för att kunna lita på den. Man måste alltså kunna spåra informationen och förstå vad som har hänt med den. I annat fall kan det bidra till något som Respondent C tar upp som ett av de största problemen; att man inte bara får en enorm mängd data, utan man har samma data överallt, alltså en sorts dålig redundans. Problemet med att kunna spåra sin data tar även Pia upp och menar att det är av stor betydelse inom bankbranschen då rapporter kring data som använts för flera år sedan kan komma att krävas när som helst. Hon fortsätter berätta att de nuvarande systemen och verktygen inte har den fulla kapacitet av spårbarhet som hon önskar.

4.3.3 Subjektiv funktionalitet

Pia på SEB tycker att deras huvudsakliga verktyg, *Informatica*, brister i att inte inkludera viss funktionalitet som de behöver inom processer för visualisering och byggandet av datamodeller, eller i detta fall *knowledge graphs*, som underlättar i ett senare skede av datahanteringen. I stället använder SEB *Tableau* som komplement för visualisering, vilket innebär en onödig kostnad, men som, enligt Pia, fungerar väldigt bra. Den största bristen på funktionalitet handlar om problematik kring utvecklandet av datamodeller, som Pia nämner som en viktig del i hur de ska kategorisera och hitta rätt data att använda. Pia fortsätter att

beskriva effekten av att inte ha den här funktionaliteten. Hon menar att det "kan gå riktigt illa" om man inom bankväsendet inte på ett bra sätt kan identifiera rätt data.

Respondent B ser bristen på funktionalitet på ett mer överskådligt plan, och nämner att molntjänster inte är lika komponent-starka som on-premiseslösningar. Vidare förklarar Respondent B att det enligt deras organisation finns fler inbyggda komponenter inom standardiserade on- premises verktyg och lösningar än molnbaserade. Det här uttalandet backar Respondent C upp med att nämna begränsningar inom *Amazon Web Services* (AWS) mikrotjänster. Respondent C nämner att de råkat på problem med *Amazon Lambda*, en mikrotjänst inom AWS, där de i början kunde lösa all funktionalitet med hjälp av mikrotjänsten men att tiden numera inte räcker till, vilket resulterar i en form av timeout.

4.4 Organisatoriska begränsningar

"Tekniken har funnits i 4–5 år men det är verksamheterna som ligger efter"
- Mats Dahl, KPMG

Mats var noga med att betona de operativa, mer organisatoriska aspekterna i Big Data management, och förklara att i och med den ständiga, exponentiella utvecklingen av tekniken och mängden data som hela tiden genereras, så hänger inte verksamheter med och klarar sig inte över den tröskeln. Detta resulterar i en gigantisk uppförsbacke för respektive verksamhet. Samtliga respondenter uttryckte oro kring detta, och nämnde några specifika områden eller begränsningar där de ansåg de nuvarande organisatoriska bristerna fanns.

4.4.1 Kostnad

Då frågan om varför respektive organisation valt att behålla sina nuvarande IT-verktyg, trots begränsningar, så nämndes i samtliga intervjuer pengar. Respondent A nämnde att allt egentligen handlar om kostnad för organisationen kontra vad de får ut av sin data eller IT-verktyg. Givet ett byte påpekar Respondent A att de skulle behöva se detta långsiktigt, och menar att om de skulle byta IT-verktyg eller system så skulle de inte se någon kortsiktig vinst, vilket kan göra dem som ansvarar för bytet negativt inställda. Detta stöds av Pia på SEB som nämner att en av de största faktorerna för att organisationer inte byter IT-verktyg som redan är etablerade, men som innehåller tydliga brister, är just kostnaden. Respondent B anser att de större IT-verktygen som löser diverse, nuvarande begränsningar är alldeles för dyra och nämner "flera miljoner" som ett pris IT-verktygen kan komma upp i. Denne avslutar med att poängtera att det kan kännas som mycket för de flesta organisationer.

För Respondent C är kostnaden en viktig faktor och denne belyser detta med exempel på ett antal begränsningar hos deras IT-verktyg. Effekten av dessa begränsningar menar Respondent C är att de alltid måste iterera genom sina nuvarande lösningar och se vad de kan ersätta IT-verktygen med i framtiden och nämner, i likhet med vad Respondent A påpekar, att problemet finns högre upp i hierarkin. Detta grundas i sin tur i att organisationens ROI (*Return On Investment*) på ett byte inte skulle vara tillräckligt högt. ROI kan alltså vara svårt att hantera på rätt sätt, menar Respondent C, eftersom en förändring inte alltid betyder en monetär vinst, men kan i längden bidra till en mängd sparad tid som i sin tur bidrar till ökade resurser. Avslutningsvis nämner Respondent C att det inte alltid är självklart hos de som betalar, och att det är därför bytena inte är så lätta att utföra.

4.4.2 Kunskap och mognadsgrad

En annan omtalad organisatorisk begränsning var enligt majoriteten av respondenterna kunskap och mognadsgrad bland medarbetarna. Respondent A nämner att de har väldigt stor "isolerad kunskap" (*single point of risk*), och förklarar att de har ett antal internt skräddarsydda IT-verktyg som en eller några få enstaka har kunskap inom. Det betyder att ifall någon ansvarig för ett visst IT-verktyg skulle sluta står företaget utan kunskapen om det verktyget eftersom de så att säga lagt alla ägg i samma korg. Respondent A fortsätter påpeka att det kan uppstå problem kring användandet av deras IT-verktyg om nya personer anställs, i och med att de har så många internt skräddarsydda applikationer och IT-verktyg. Detta är ingen begränsning för företag med standardiserade IT-verktyg, menar Respondent A, eftersom de IT-verktygen alltid brukar vara väldokumenterade och användarvänliga, vilket

deras egna inte är. Som skäl till detta nämner Respondent A svårigheten att utbilda folk inom de egna processerna och IT-verktygen. Denna begränsning betonar även Pia, och förklarar att det gällande kunskapsbristen handlar om både ledning, medarbetare och kund. Alla pratar om tekniken, ingen pratar om kunskapen eller mognadsgraden, säger Pia, och adderar att de på SEB har tilldelade yrkesroller som respektive medarbetare inte vet vad de innebär, och menar vidare att de nuvarande arbetsuppgifterna går före nya direktiv från ledningen gällande dessa roller.

Respondent B nämner en generellt låg mognadsgrad, och förklarar att organisationer inte vet om att de kan få Big Data-lösningar att fungera, eller vad de ska göra med den, och att de saknar långsiktiga Big Data-planer. Respondent B ger ledningen inom organisationerna skulden för den låga mognadsgraden. Hen upplyser om att större företag ofta styrs av ekonomer som i princip alltid blickar bakåt till hur man tidigare gjort i hopp om att investera i nya lösningar som kan ge ett ekonomiskt värde för företaget. Istället borde man, enligt Respondent B, se över problemen i de nuvarande IT-verktygen eller lösningarna och försöka göra dem bättre för att effektivisera verksamheten. Detta ser Mats på KPMG också som en större svårighet; att få ledning, experter och individer som inte jobbar så mycket inom IT att förstå hur viktigt det är att fatta korrekta beslut inom ämnet. Mats förklarar att ledningsgruppen måste förstå att det är en väldigt strategisk fråga angående kunskap om IT generellt. De måste komma överens om hur de ska se på, hantera och interagera sin data för att bygga applikationer, med eller utan hjälp av IT-verktygen, som ger värde för kunden. Detta för att i slutändan använda den datan för att dra slutsatser inom andra delar av deras verksamhet. Mats menar alltså att ledningen måste komma över en tröskel och förstå vad tekniken, med ändrade affärsmodeller och sätt att bedriva verksamheten på, kan innebära för respektive organisation. En ytterligare, mindre uppmärksammad organisatorisk brist nämndes av Pia och handlar om användarvänligheten bland vissa av IT-verktygen som kunderna hos SEB använder. I och med att det inte är några utbildade, tekniska personer som använder sig av dessa IT-verktyg så menar Pia att det kan uppstå svårigheter och brist på förståelse kring dem, vilket i sin tur kan orsaka problem längre fram i processen.

4.4.3 Volymhantering

Respondent A tog upp frågan om volymhantering inom deras processer, och ifrågasatte hur de på ett optimalt sätt ska veta vilken och hur mycket data de vill spara. Att spara information eller data som inte används och bara tar upp lagringsutrymme menar Respondent A bara är en onödig kostnad, men poängterar samtidigt utmaningen att veta att just den datan som man då möjligtvis raderar inte skulle, i framtiden, komma till användning. Respondent A fortsätter att exemplifiera problemet och tar upp att det är omöjligt att spara alla data för dess potentiella värde, i och med kostnad och lagringsutrymme.

“Hur kommer det se ut om vi sparar all data som vi genererar?” - Respondent A

För närvarande upplever Respondent A att de inte genererar eller använder så stora mängder data att de behöver standardiserade lösningar som *Microsoft Azure* eller *Amazon Web Services*, men är noggrann med att det främst är Big Data de jobbar med. Respondent A fortsätter förklara att det enbart är internt utvecklade verktyg och system de jobbar med, men poängterar även att bytet till en mer standardiserad lösning gällande deras verktyg är given om behov att behandla större datavolymer skulle uppstå.

4.4.4 Förändringsarbete

Att vara öppen för förändring är något majoriteten av våra respondenter värdesätter, men detta kan vara lättare sagt än gjort inom en större organisation med många anställda. Respondent B nämner att deras anställda helst inte vill skruva på något som redan fungerar. Vidare nämner Respondent B att ledning och medarbetare tycker att processer och IT-verktyg fungerar bra, och de frågar sig själva varför man ska ändra något som redan fungerar, men tänker inte på att något kan fungera bättre, mer effektivt. Det här nämner även Pia som påstår att tekniken aldrig varit ett stort problem, utan det är svårare med organisation och kultur, framförallt att få igång ett fungerande förändringsarbete.

Som tidigare nämnts tog Mats på KPMG upp en sorts kunskapströskel att komma över för att förstå hur tekniken kan gynna ens organisation. Givet förändringsarbetet så fortsätter Mats berätta kring detta, och poängterar att förändringsarbete är en stor organisatorisk begränsning vilken hindrar möjligheten att optimera processer inom Big Data management. Mats understryker detta med egna erfarenheter och påtalar att han jobbat länge inom IT-industrin och haft egna förutfattade meningar vilket hindrat honom att ta rätt beslut i vissa situationer, som enligt honom hade gynnat organisationen markant.

4.4.5 Data Governance

En annan aspekt kommer från Pia som berättar hur de på SEB lägger stor vikt vid deras *Data Governance*-program för att försöka höja sin datakvalitet och skapa regler kring hur de använder sin heterogena data. Hon nämner dock också en form av begränsning i detta när de först och främst inte vet vilka data de ska jobba med, och inte heller inser svårigheten av att hämta fram sin data ur sin Data Lake. Respondent B tar också upp dubbeljobb vid talan om kvalitetsbrist, och pekar på hur ineffektiv deras upprätthållande av datakvalitet är och att det blir tydligt dubbeljobb eftersom systemen inte kan hantera det själva. Respondent B nämner ordagrant att det är inom källsystemen som denna begränsning oftast ligger, vilket resulterar i att verksamheterna och dess medarbetare får rätta fel i efterhand. Detta bidrar till dubbelarbete och en ökad kostnad för organisationen. Avslutningsvis menar Respondent C att hantering av information för anställda i sig egentligen inte är något större problem, utan problematiken ligger i hur man ser till att hantera rätt information; definiera, identifiera och skala bort känslig information, och fastställa att rätt information är tillgängligt för rätt individer.

4.5 Sammanfattning av resultat

Utifrån ovanstående identifierade begränsningar från respektive intervju har vi sammansatt en tabell som förklarar de aspekter som begränsningarna ingår i, och även för vilka respondenter begränsningen var aktuell.

Tabell 5: Sammanfattade begränsningar utifrån teknisk eller organisatorisk aspekt

Aspekt	Begränsning	Beskrivning	Respondenter
Teknisk	Kompatibilitet	Svårigheter att byta ut verktyg då organisationer vill bibehålla kompatibiliteten mellan dessa	Respondent A, Pia, Mats
Teknisk	Spårbarhet	Brist på spårbarhet innebär en minskad tillförlitlighet kring ens data och åtkomst av historik	Pia, Respondent B, Respondent C
Teknisk	Datakvalitet	Dålig grund att bygga på	Respondent A
Teknisk	Subjektiv funktionalitet	Funktionaliteter inom använda IT-verktyg som organisationer saknar och hittar supplement till	Pia, Respondent B, Respondent C
Organisatorisk	Kostnad	Motsätter värdeskapande	Respondent A, Pia, Respondent B, Respondent C, Mats
Organisatorisk	Kunskap och mognadsgrad	Fundamentalt problem inom organisationer som sätter grunden till flera andra begränsningar	Respondent A, Pia, Respondent B, Respondent C, Mats
Organisatorisk	Volymhantering	Kännedom inom organisationerna att veta vilka mängder data att spara som ska ge värde för organisationen	Respondent A
Organisatorisk	Förändringsarbete	Limiterar möjligheten att effektivisera delar av datahanteringen	Respondent B, Respondent C, Pia, Mats
Organisatorisk	Data Governance	Problematik kring regelsättning av heterogena data och kännedom kring vilken data de ska använda	Pia, Respondent B

5 Analys

I den här delen kommer vi analysera våra resultat och dra kopplingar till vår litteraturgenomgång om sådana finns. Vi kommer se över vad respondenterna upplyser och se ifall det finns likheter eller skillnader gentemot redan etablerad teori.

5.1 Definitioner

5.1.1 Big Data

Majoriteten av våra respondenter anser att Big Data främst handlar om stora datavolymer, vilket går i linje med vår litteraturgenomgång och vad Raheem (2019) anser, som även påvisar att det handlar om tre större strukturer; strukturerad, semistrukturerad och ostrukturerade data, som Respondent C nämner i sin definition av Big Data. Respondenterna definierar även Big Data som arbetssätt som involverar hantering av detta, men har egentligen olika definitioner som helhet. Mats definition är mer teknisk och ur ett optimeringsperspektiv, vilket är intressant att poängtera och hänvisar tillbaka till vad Raheem (2019) nämner angående vikten av vad man gör med datan, och inte enbart att det är volymen som är det viktiga. Vi ser även ett optimeringsperspektiv som extra intressant eftersom ett av våra syften är att hitta utrymmen att hjälpa kring att optimera processer för Big Data. Respondent C nämner 5 V-modellen när hen pratar om Big Data, och poängterar att hen är kluven gällande modellen, vilket är en intressant iakttagelse eftersom det går i linje med vad vi tidigare nämnt, och vad som presenterats av Ferguson (2012), angående flertalet olika V-modeller för definiering av Big Data. Pia nämner att Big Data är möjligheten att när som helst kunna skala upp och ner, vilket backar upp det Davenport (2012) nämner om att Big Data inte enbart behöver definieras som stora mängder data, utan att man även kan beakta skillnader på vanliga data och Big Data.

Som helhet anser vi det intressant med olika definitioner av Big Data. Dessa definitioner kan, enligt oss, ses ur ett perspektiv där definitionerna bland användare och utvecklare möjligen kan resultera i att IT-verktyg används eller utvecklas med olika mål i sikte, eller utifrån olika perspektiv, vilket skulle kunna bidra med problem beroende på vem som använder IT-verktygen.

5.1.2 Begränsning

Samtliga av respondenternas definition av en begränsning går i linje med hur vi definierar det i Litteraturgenomgången. Respondent A nämner att det även handlar om att man väljer ett annat IT-verktyg som innehar den funktionaliteten, vilket väcker en intressant fråga gällande resurser kring byte av IT-verktyg inom organisationerna. Vi nämner i litteraturgenomgången att begränsningar som verksamma inom Big Data upplever är en effekt av att IT-verktyg inte löst ett visst problem inom organisationen, vilket tydligt speglas i respondenternas definitioner. Pia exemplifierar detta i sin definition och menar att en begränsning är att ett verktyg inte klarar av att utföra en viss funktionalitet som man förväntar sig, och att man då provar andra verktyg för att klara att kringgå problemet eller utmaningen.

5.2 Tekniska begränsningar

5.2.1 Kompatibilitet

Kompatibilitet är något som anses viktigt från både Pia, Respondent A och Mats sida. Pia och Respondent A påtalar i princip samma saker när de berättar om kompatibiliteten hos deras implementerade IT-verktyg för Big Data management. Båda använder fortfarande en hel del *legacy system* som just nu ligger utspridda för att sköta olika saker. Pia berättar att man har upp emot 40 sådana system som man just nu jobbar med att integrera till ett enat system. För att detta ska bli verklighet måste filtyper och applikationer inom systemen bli kompatibla. Respondent A berättar att deras *legacy system* besitter dålig kompatibilitet gentemot både internt utvecklade verktyg och script och mot nyare verktyg. Hen påpekar även att det är svårt att uppdatera dessa system eftersom man riskerar att försämra kompatibilitet med andra verktyg. Mats anser också att deras verktyg måste bli mer kompatibla för att kunna hantera flera olika typer av data, oberoende vilken kund man jobbar med just då. Att uppleva begränsningar i kompatibiliteten mellan sina implementerade verktyg och system för att hantera Big Data ger, enligt oss, också sämre förutsättningar för att kunna integrera processade data i organisationen. Att integrera data är ett känt problem inom Big Data management där aggregering och kategorisering kan leda till problem (Chen et al., 2013; Labrinidis et al., 2012; Sivarajah et al., 2017).

Av de respondenter som påtalat bristande kompatibilitet mellan system och verktyg så använder två av dessa till större del skräddarsydda lösningar och *legacy system.* Detta beror på att organisationer i ett tidigare skede har byggt dessa system och inte insett, eller inte fått medhåll från ledningen; att kunna integrera systemen och/eller byta till standardiserade lösningar. Detta kräver att samtliga filtyper och applikationer inom dessa system måste vara kompatibla med varandra för att fungera ihop i ett aggregerat system, framför allt vid förändring av verktygen och ifall man av någon anledning skulle behöva byta verktyg. Detta är så klart en teknisk begränsning och något man borde funderat över när man utvecklade systemen från början. Detta har sin grund i en okunskap hos ledningen, där de inte kan se det bestående värdet i att byta till en standardiserad lösning eller bygga om för att förbättra kompatibilitet. Om organisationer inte eftersträvar bättre kompatibilitet kommer det sannolikt inte att hålla i längden, i takt med att datamängderna ökar och "bäst före" minskar. Man måste kunna aggregera och integrera data snabbare för att faktiskt kunna få ut något värde av det.

5.2.2 Spårbarhet

Tre av dem vi intervjuade uttryckte att spårbarheten inom deras verktyg var bristfällig. Respondent B berättar att man använder Microsofts standardiserade lösningar och IT-verktyg, där det är relativt lätt att följa sin data, för att hantera Big Data, men att det inom deras *Oracle* databaser och *DB2* finns påtagliga brister i spårbarheten av datan. Även *Informatica* påstås ha brister på denna punkt, enligt Pia. Respondent C fyller på och menar att spårbarhet är det största problemet inom deras organisation och dess IT-verktyg för att hantera Big Data. Hen nämner också att det är mycket viktigt att kunna spåra vad som har skett med datan för att kunna lita på den och förstå varför den ser ut som den gör i ett visst skede av processen. Detta är något som Pia också är beroende av, och hon förklarar att man måste kunna spåra sin data långt bak i tiden för att med säkerhet kunna uppfylla de krav som finns inom banksektorn när det gäller olika typer av rapporter.

Detta fenomen kan spåras tillbaka till att man innan analys måste ha tillförlitlighet till datan (Gantz et al., 2012; Sivarajah et al., 2017). Tillförlitlighet till datan ges, enligt Respondent C, genom bland annat god spårbarhet och transparens, lite som en KRAV-märkning inom datahantering. Spårbarhet på datan är något som, enligt oss, också är viktigt för att kunna felsöka potentiellt defekt information. Ju fler avdelningar inom en organisation som får tillgång, desto fler kan sannolikt transformera datan för deras specifika behov. I detta skede kan defekter antagligen uppstå och då är det viktigt att kunna spåra datans bana genom organisationen för att kunna fastslå varför den är defekt.

5.2.3 Subjektiv funktionalitet

Pia inleder detta stycke med att påpeka bristfällighet i deras verktyg *Informatica.* Bristen upplevs under visualiseringen av datan och byggandet av datamodeller, i detta fall kallade *knowledge graphs*. Det har lett till att man implementerat *Tableau* som komplement för visualiseringen vilket, enligt Pia, fungerar bra men är en onödig kostnad. Huvudproblemet ligger dock i sammanställandet av deras *knowledge graphs* som är viktig för deras kategorisering och lokalisering av rätt data för rätt användningsområde. Pia beskriver också effekten av dessa begränsningar som mycket påtagligt då det är viktigt att kunna identifiera rätt data inom bankväsendet. Respondent B berättar mer överskådligt att det, enligt deras organisation, återfinns bättre funktionalitet inom on-prem lösningar i jämförelse med molnlösningar. Härnäst förklarar Respondent C att man upplever begränsningar inom *Amazon Web Services (AWS)* mikrotjänster. Specifikt talar denne om *Amazon Lambda* som tidigare enligt Respondent C definierats som mindre funktioner vilka bara kan exekveras under en viss tidsrymd. Detta fungerade ett tag men Respondent C hävdar nu att de inte längre kan utföra de tänka uppgifter som man använder *Amazon Lambda* till inom samma tidsrymd vilket leder till en form av timeout.

Det Pia pratar om kan direkt kopplas till litteraturen där vi tar upp att visualisering av data kan vara svårt eftersom Big Data generellt består av ett antal olika dimensioner som gör den svårhanterad (Sivarajah et al., 2017). Själva problemet ligger i att man fått använda komplement till redan implementerade verktyg för att skapa god visualisering av datan. Pia har tidigare också berättat att man har upp till 40 olika system som sköter datahanteringen, och ovan nämnda problem skulle vi säga är en effekt av detta. Man har inte särskilt bra kompatibilitet eller integrering mellan systemen. Därför är det svårt, när man upptäcker en brist såsom visualisering, att byta ut verktyget mot något bättre eftersom systemarkitekturen antagligen är väldigt instabil och beroende av sin omgivning. Respondent C upplever begränsningar med en mikrotjänst inom *AWS.* Organisationen har redan från början räknat med att man i ett framtida scenario kan få problem med Lambda, men eftersom *AWS* fyller deras andra krav har de överseende med denna begränsning. Detta är en ren brist i verktyget som de använder sig av och eftersom man inte upplever övriga, tekniska problem med plattformen byter man sannolikt inte och får rimligtvis förlita sig på att Amazon uppdaterar Lambda i framtiden.

5.3 Organisatoriska begränsningar

5.3.1 Kostnad

Kostnadsaspekterna dök upp som en återkommande utmaning för de organisationer som vi intervjuade, framförallt i kombination med frågor gällande byte av nuvarande IT-verktyg. Respondent A nämnde att kostnaden egentligen är allt det handlar om; och poängterar att kostnaderna måste vara lägre än det värdet man får ut av användningen av sin data. Hen fortsätter att berätta hur ledningen lätt motsätter byte av IT-verktyg på grund av en kortsiktig kostnad, där de inte ser den långsiktiga vinsten. Detta går i linje med vad Al Nuaimi et al. (2015) konstaterar gällande att höga kostnader, kring implementering av nya lösningar och IT-verktyg, kan sätta käppar i hjulet för effektiviseringen kring sin databehandling. Detta stödjer Pia i sitt uttalande om att den största faktorn för att organisationer inte byter IT-verktyg är just kostnaden. Respondent B hävdar att IT-verktyg med omfattande funktionalitet kan komma upp i miljonbelopp, vilket kan kännas dyrt för de flesta organisationer. Detta stämmer överens med vad Sivarajah et al. (2017) påvisar angående dyra strategier för att effektivisera komplexa mängder data, och att det krävs akuta strategier för att kraftfullt minska kostnaderna eftersom IT-verktygen inom dessa strategier idag är alldeles för dyra. Likväl kopplar det tillbaka till att skilja på att hantera vanliga data och Big Data i valet av metoder och IT-verktyg, och även höga varianter på ens Big Data (Almeida, 2017; Chen et al., 2013; Labrinidis et al., 2012; Vaghela, 2018), som bidrar med ännu ett argument som kan kosta organisationer stora av pengar (Sivarajah et al., 2017). Pia poängterar även problemen kring kostnader gällande datakvalitéer hos deras *Data Governance*-program, vilket speglar det Sivarajah et al. (2017) och Almeida (2017) menar med utmaningen att säkerställa, genom kravet på IT-verktyg, höga datakvalitéer efter att extrahering och lagring är utförda. Tidigare fastställde vi hur Sivarajah et al. (2017) ser på att kostnaderna alltid kommer vara en stor utmaning inom Big Data management i och med ständigt ökade heterogena dataströmmar som extraheras och genereras av organisationer. Detta backas även upp av Al Nuaimi et al. (2015), vilket Pia gav prov på med att de på SEB hade svårigheter att veta vilka typer av data de behöver hantera. Detta visar en problematik gällande kostnader och även om kunskap och mognadsgrad medarbetare och ledning inom respektive organisationer.

Att ledningen tycks motsätta sig hos byten av IT-verktyg på grund av brist på kortsiktig vinst ger oss en känsla av avsaknad av kunskap hos ledningen, att ledningen bör känna till att inget direkt värde kommer att synas men att man måste blicka framåt och se hur bytet kan gynna organisationen långsiktigt. Här kan vi även konstatera att det krävs billigare supplement till omfattande IT-verktyg och lösningar så att alla har råd med en effektiv hantering av Big Data. IT-verktyg kräver idag möjlighet att kunna ta hand om stora mängder heterogena data, men i och med utvecklingen och ökningen av Big Data så kan man konstatera att mängden typer av data också kommer att öka, vilket sätter press på IT-verktygen. Överlag är det lätt att konstatera att majoriteten begränsningar som kan uppstå utifrån problem och utmaningar kan kräva stora resurser för att lösa, vilket inte direkt är någon större nyhet, men som strategier kommer att behöva utvecklas runt för att minimera.

5.3.2 Kunskap och mognadsgrad

Respondent A nämner att de förföljs av isolerad kunskap, alltså ”*a single point of risk*”, och förklarar att de har internt skräddarsydda IT-verktyg med fåtalet anställda med kunskap om. Detta är en intressant iakttagelse som vi anser är mer vanlig inom internt utvecklade IT-

verktyg snarare än standardiserade lösningar, i och med den lättillgängliga och tydliga dokumentationen för IT-verktyg som brukar finnas inom större standardiserade lösningar. Al Nuaimi et al. (2015) tar upp utmaningen med att upprätthålla kunskap kring Big Data-hantering bland medarbetare. I och med en isolerad kunskap så riskerar företaget att förlora kunskapen om IT-verktyg och applikationer för gott, om anställda som har kunskapen lämnar organisationen. Pia betonar detta och nämner att alla pratar om tekniken, men ingen pratar om kunskap eller mognadsgrad. Sivarajah et al. (2017) lyfter fram vikten av att implementera infrastruktur inom organisationer för hantering och bearbetning av Big Data, och menar att detta kräver en nivå av kunskap hos medarbetare för att förstå sig på Big Data och användningen kring det. Vidare förklarar Respondent A att det är svårigheter kring utbildning som ligger till grund för den isolerade kunskapen, i och med de internt utvecklade IT-verktygen. Det här går i linje med vad vi tidigare ansåg utgöra problem kring internt utvecklade lösningar och bristen på dokumentation. Respondent A poängterar vidare att det ofta inte är någon begränsning för större företag med standardiserade lösningar och IT-verktyg just på grund av att de är väldokumenterade och användarvänliga, vilket deras interna IT-verktyg och applikationer inte är. Den här typen av brist känns väldigt abstrakt, och är inget man enkelt kan fastställa utan att ha varit praktiskt involverad i processerna kring Big Data management inom respektive organisation. Respondent B följer samma fotspår som Pia och menar att det finns en generell omognad inom organisationer kring hantering av Big Data. Här ges ledningen återigen skulden, vilket visar att det är ett stort problem och en genomgående begränsning inom organisationen som helhet. Mats tar upp något intressant gällande förståelse kring generell IT hos ledningen, vilket vi anser vara en fundamental kunskap om man ska driva, eller sitta i ledningen, för ett större IT-bolag. Mats fortsätter att betona hur viktigt det är med kunskap inom valfritt ämne för att fatta korrekta beslut Det känns som om organisationer bara rullar på, tekniken utvecklas exponentiellt, men att ledningen saknar kunskapen för att kunna fatta korrekta beslut i samma takt som utvecklingen sker. I längden kan detta begränsa organisationen, och möjligtvis resultera i att man tappar mark gentemot konkurrenter.

5.3.3 Volymhantering

Respondent A nämner svårigheter med att veta hur mycket data man ska spara. Kostnaden att lagra data som inte används känns onödig. Svårigheter att optimera processen; att veta exakt vilka mängder man ska spara för att använda, så att ingenting går förlorat. Chen et al. (2013) uttryckte en utmaning för IT-verktyg att göra datan snabbt tillgänglig för organisationer, vilket även backas upp av Adiba et al. (2016) och Vaghela (2018), och behöver överkommas om problemet Respondent A tar upp skulle ha chansen till att lösas. Vidare nämner Respondent A att de skulle behöva gå över till mer standardiserade lösningar, om behovet av att behandla större mängder datavolym uppstod, vilket går i linje med problematiken kring höga volymer data som Sivarajah et al. (2017) tar upp gällande krav på nya metoder och IT-verktyg.

När det gäller svårigheter med att veta vilka optimala mängder data som organisationer ska spara och lagra, så finns idag ingen lösning. Det vore välkommet för företag om de enkelt kunde veta exakt, innan de extraherar och lagrar datan, vilken data som skulle gynna organisationen mest och ge högst värde. I nuläget känns detta omöjligt med den tekniken och de IT-verktyg som finns tillgängliga. Dock känns problemet vanligare inom organisationer som genererar sin egen data, men framstår fortfarande som ett subjektivt problem bland företag, beroende på vilken sorts data de samlar in.

5.3.4 Förändringsarbete

Respondent B nämner att medarbetare, så väl som ledning, anser att när väl IT-verktyg och viss funktionalitet fungerar bra, så vill man inte skruva på något som redan fungerar. Det här går i linje med vad Russom (2019) påstår kring att anpassa sig till nya designparadigmer. Respondent B, och även Pia, nämner att ledningen menar "Varför ändra på ett vinnande koncept?", och ser bara en kortsiktig kostnad och brist på inkomst snarare än en långsiktig förändring och värdeökning. Det handlar givetvis om flertalet människor som ska jobba tillsammans i symbios för att ett förändringsarbete ska gå att genomföra smärtfritt, vilket stämmer överens med vad Russom (2019) och Sivarajah et al. (2017) påstår kring detta organisatoriska problem. Vidare menar Russom (2019) att anpassning för förändring är ett steg i rätt riktning för organisationer att gå från en suboptimal Big Data-hantering till en bättre. Mats påpekar en form av kunskapströskel gällande förändringsarbete, som backar upp vår tanke kring organisatoriska begränsningar, att det kan begränsa möjligheterna att utveckla och i sin tur optimera processer inom Big Data management. Detta är extremt viktigt för den interna utvecklingen inom processer och användningen av IT-verktyg för organisationer, men där tillvägagångssättet för ett bättre organisatoriskt förändringsarbete känns väldigt subjektiv beroende på bransch och företag. Den vy på förändringsarbete som våra respondenter speglar anser vi visar att organisatoriska begränsningar inom organisationer kan sätta stopp för den tekniska aspekten. Detta kopplar även tillbaka till avsnittet om Kunskap och mognadsgrad, där vi diskuterade vikten av att ha kunskap om vad förändring kan göra för organisationen i längden. Om ledningen inte bygger upp en förståelse kring vad en förändring, stor som liten, kan göra för organisationen så sätter det käppar i hjulet för innovation och effektivisering kring respektive process eller IT-verktyg.

5.3.5 Data Governance

Pia nämner jobbet inom deras *Data Governance* program som väldigt viktigt för att säkerställa god struktur kring regler, användning och kvalitéer av sin data och hur de lägger stor vikt vid i det. Vidare förklarar Pia problematiken, i form av en begränsning, i hur de vet vilken data de ska jobba med utifrån deras Data Lake. Den här begränsningen går hand i hand med vad Almeida (2017) och Sivarajah et al. (2017) nämner om problem med att hämta data från organisationers insamlade pol. Pia anser att det blir en form av dubbeljobb som både blir kostsamt och tidskrävande, vilket påvisar vikten av att utveckla nya procedurer att utvinna datan på som Chen et al. (2013) nämner kan komma att leda till högre effektivitet och komma närmare optimal datahantering. Vidare belyser Respondent B en form av dubbeljobb gällande fastställande av datakvalitet, och påpekar att deras system själv inte kan hantera det. Detta går i linje med Labrinidis et al. (2012) som betonar kraven på system och IT-verktyg; att kunna hantera mängder heterogena data, vilket i annat fall kan fördröja processer och minska effektiviserad Big Data management som helhet inom organisationer. Respondent B menar att detta i sin tur bidrar till en ökad kostnad för organisationer, att delvis behöva hitta alternativa lösningar till problemet eller att processerna helt enkelt tar för lång tid att utföra, vilket Sivarajah et al. (2017) menar kan bidra till ökad kostnad ifall nya system eller IT-verktyg som kan hantera detta behöver köpas in.

Det här känns som ett problem som är väldigt svårt att lösa, där korrigeringar måste göras i ett tidigt stadium vid extrahering av data. Ifall man inte har något syfte med på den data man extraherar så kommer man ha större mängd data insamlad som man måste gå igenom för att hitta rätt data, som ger värde för organisationen. Vi anser att större reglering och tanke måste läggas på den data som man extraherar. En annan synvinkel som motsäger det vi just nämnde

är kännedomen kring vilken data som då skulle extraheras och inte, vilket för oss tillbaka till avsnittet om Volymhantering. Detta resulterar i en form av paradox där man både måste veta vilken data man vill extrahera för att minimera onödig lagring, men samtidigt först extrahera data för att se vilken data som är värd att spara.

6 Slutsats

En effektiv hantering av Big Data kräver av en organisation att både de tekniska och organisatoriska aspekterna fungerar. Detta ställer krav på organisationer att delvis bibehålla sin flexibilitet och att ha tillgång till IT-verktyg och system som kan hantera de otroliga mängderna heterogena data som genereras/extraheras och används. Även att medarbetare och ledningsstab besitter den kunskap, intresse och vilja som krävs för att kunna arbeta och på ett effektivt sätt optimera arbetssätt, användningen av IT-verktyg och system inom dessa processer.

Syftet med vår studie är att lyfta fram vilka begränsningar som verksamma inom Big Data management finner hos användningen av IT-verktygen, och identifiera möjliga bakomliggande faktorerna kring dessa. Vi har under studiens gång sett problemet ur en teknisk synvinkel och utgått ifrån att det är inom de tekniska delarna av IT-verktygen som begränsningarna finns, men att det i själva verket mer handlar om begränsningar ur en organisatorisk eller operativ synvinkel. Tekniken finns där, men man måste kunna förstå sig på, vara anpassningsbar och jobba med tekniken på rätt sätt, vilket ligger som grund för att processer för Big Data management inte effektiviseras optimalt.

Begränsningar eller problem som vi tar upp i teorin, och som även identifieras i resultatet, behöver nödvändigtvis inte lösas av funktionella IT-verktyg. Identifierat från respondenternas svar är organisatoriska begränsningar som kostnad, kunskap och ställningstagande mot förändringsarbete de största bovarna och anledningen till att organisationer inte effektiviserar sina processer inom hantering av Big Data, men även volymhantering och Data Governance, och tekniska begränsningar som spårbarhet, kompatibilitet och datakvalitet. Det är kostnaden, bekvämligheten för, och kunskapen kring IT-verktygen som är de största problemen, vilka grundas i att ledning, såväl som medarbetare, saknar intresse och kunskap om hur man ska ställa sig till effektivisering av Big Data management när allt utåt verkar fungera som det ska.

Organisationer måste, på ett annat sätt än idag, vara villiga att jobba mot utmaningar och begränsningar. Det vi har sett är att kunskapen kring vad som är viktigt för optimal hantering av Big Data i första hand ligger hos de som jobbar med det, vilket är rimligt. Problemet är att dessa kunskaper inte representeras på rätt sätt högre upp i hierarkin och därför är det svårt att uppnå den nivån av kunskap och expertis runt användandet av IT-verktygen som kan ge det värde som man eftersträvar.

6.1 Vidare forskning

Vi anser att vår metodik genom hela studien, och det resultat vi fått fram genom användning av denna, besvarar vår forskningsfråga i en större omfattning. Med resultatet av studien i åtanke finns det dock utrymme för vidare forskning inom ämnet. Läsaren bör förstå att studien är baserad på ett dataset om fem (5) respondenter eller organisationer. För att få ett noggrannare och mer genomgående resultat bör man skala upp studien och samla in empirisk information från flertalet organisationer och respondenter. Detta skulle hypotetiskt sett visa en högre avsaknad av subjektiva funktionaliteter, och även ett högre antal begränsningar som de verksamma upplever, samt stärka argumenten för höga kostnader, brist på kunskap, och dåligt

ställningstagande mot förändringsarbete inom organisationerna. Man kan även avgränsa sig till en specifik bransch eller begränsning för att ge undersökningen ett djupare syfte.

Vi har utgått ifrån IT-verktygen som hanterar Big Data för att hitta begränsningar inom dessa, där slutsatsen resulterade i att det finns ett antal tekniska begränsningar men att de flesta av dessa baseras utifrån organisatoriska aspekter. Förslagsvis kan man bygga på vår studie och undersöka hur IT-representeras högst upp i beslutsfattandet inom organisationer. Vår studie indikerar att det just nu inte riktigt fungerar på en ledningsnivå när det kommer till att fatta beslut som gynnar hantering av Big Data. Det är, dock, just en indikation och det är därför intressant att studera detta i ett större spektrum på ett större dataset för att få fram tydligare resultat.

Appendix 1: Intervjuguide

Del 0: Introduktion:

- Introduktion av intervjuare (namn, skola, utbildning)
- Förklara målet med studien, vår definition av Big Data och hur lång tid intervjun är

1. Har vi ditt tillstånd att transkribera den här intervjun och inkludera den i vår studie?
    a. Om ja, signalera att vi börjat transkribera.
2. Vill du förbli anonym eller är du okej med att vi hänvisar till dig i vår studie?

Del 1: Bakgrund och Big Data management:

1. Vem är du och vad innebär din arbetsroll?
2. Hur länge har du jobbat inom det nuvarande företaget?
3. Hur skulle du definiera "Big Data"?
4. Förklara en översikt av er allmänna datahantering för Big Data
    a. Skapar ni er egen data eller hämtar ni från tredje part?
    b. Hur ser de olika processerna (eller stadierna) i er datahantering ut?
        i. Hur fungerar de?

Del 2: IT-verktyg:

1. Berätta för oss vilka IT-verktyg ni använder inom era processer för Big Data management.
    a. Hur är de utvecklade?
    b. Beskriv om de är från tredje part eller interna.
    c. Varför dessa IT-verktyg jämfört med andra?
2. Inom vilka specifika steg (eller faser) inom Big Data management använder ni dessa verktyg?
3. Känn dig fri att rangordna varje verktyg i vilken grad (1–5) dina verktyg ger din datahantering det stöd du behöver (1 är mycket låg, 5 är mycket hög).

Del 3: Begränsningar:

1. Hur skulle du definiera en "begränsning" inom en mjukvara?
2. Beskriv eventuella begränsningar inom era respektive IT-verktyg
    a. Vad är effekten av dessa begränsningar på organisationen?
    b. Vad hindrar din organisation från att byta verktyg med tanke på deras begränsningar?
3. Hur tror du att ett plötsligt behov av att hantera en ökad mängd data skulle påverka användningen av era respektive IT-verktyg?
4. Hur skulle ni ställa er mot användning av respektive IT-verktyg ifall kravet på en ökad förväntan på datakvalitet uppstod?

Del 4: Framtiden:

1. Vilka är de kortsiktiga utmaningarna för er Big Data management?
2. På vilket sätt tror du att begränsningarna av de presenterade IT-verktygen kan påverka er på lång sikt?
3. Är det något du vill tillägga som du tycker att vi missat att ta upp?

Del 5: Avslutning:

- Signal om att transkriptionen är slut.

Appendix 2: Intervju 1

Intervjuare: Christian Dahlberg (CD)
Respondent: Respondent A (Anonym)
Organisation: Företag A (Anonym)
Datum, tid och plats: 15 april 2020, 20.30, Lund (Sverige)
Intervjutyp: Digital intervju över Zoom
Tid: 37:18

[00:00]	**Introduktion**

Intervjuare (CD): **Okej, men då sätter vi igång på en gång då! Vill du bevaras anonym eller godkänner du att vi refererar till dig i vår studie?**

Respondent: Absolut, 100% anonym!

Intervjuare (CD): **Yes, och din organisation också då antar jag?**

Respondent: Ja precis. På grund av lagen i det här landet måste både jag och företaget jag jobbar för bevaras anonyma, tyvärr.

Intervjuare (CD): **Det fungerar alldeles utmärkt!**

[00:45]	**Bakgrund och Big Data management**

Intervjuare (CD): **Då sätter vi igång! Berätta lite om dig själv och om din arbetsroll.**

Respondent: Jag heter Respondent A och jag är R&D ingenjör. Jag jobbar kort sagt inom R&D (*Research & Development*) och forskar på att utveckla ny teknik inom filmbranschen. Vi utvecklar främst teknik som gör att det går... eller det handlar mest om resursbesparingar i slutändan. Vi vet redan hur man gör de flesta sakerna, får till de flesta effekterna osv, men det tar mycket beräkningskraft vilket gör att vi vill få det att gå snabbare. Allteftersom allt blir snabbare så kan man börja använda dyrare metoder som man inte har kunnat använda förut och så fortsätter det sådär. Börjar med snyggare och långsamma processer, men så försöker vi då göra den processen snabbare.

Intervjuare (CD): **Yes, okej. Och hur länge har du jobbat för Företag A?**

Respondent: Två år.

Intervjuare (CD): **Okej! Då ska vi se. Hur skulle du definiera 'Big Data'?**

Respondent: Jo, men Big Data är ju att man börjar… Det är egentligen en vidareutveckling på… Alltså, i grund och botten kommer det från statistik; att man exempelvis har någon data på försäljningssiffror, så vill

man kunna förutspå hur vissa parametrar fungerar och vad de har för inverkan på sina försäljningssiffror eller vad det nu är man tittar på. Man har en mängd data alltså, och så försöker man använda den för att förutspå saker. Men Big Data är... I modern tid har det ju kommit så extremt mycket data, och man har inte kunnat bearbeta och dra nytta av sin data för att kunna göra förutspå saker förut. Så Big Data är alltså metoder för att behandla enorma mängder data helt enkelt, för att kunna förutspå saker på ett bättre, mer skräddarsytt sätt.

Intervjuare (CD): **Ja, precis. Det låter bra. Då ska vi se… Nästa fråga då, förklara gärna en översikt på er allmänna datahantering. Exempelvis från inhämtning av datan till användning av datan.**

Respondent: Ja, det finns två olika saker jag kan komma på då som är grund till hela vår data pipeline. Antingen är det sådan vi får in data från filminspelning, vilket kan vara; filmat material, vanligt filmfoto, mätningar över ljus och sånt, data för omgivningen för alla scener så man kan ljussätta på ett lämpligt sätt osv. Så det är som en sorts *Data Ingestion* som vi då får från inspelning av filmerna som vi använder oss av. Vi kan också få dataströmmar från tredje part bolag, sådana som gjort för-visualisering på hur scenerna ska se ut. Det kan vara 3D modeller och sådant som andra företag har gjort åt oss. Alltså att vi har outsourcat dessa delar så vi kan ta in det sen. Vi kan ta över projekt från andra företag, där vi ibland får allt dom redan har gjort på disk, Så det är väl de två största sätten vi tar in data på.

Intervjuare (CD): **Yes, okej. Nu nämnde du två sätt som ni hämtar in data på. Hur ser resterande delar av datahanteringen ut?**

Respondent: Ja, precis, jag var inte riktigt färdig än. Den mesta datan vi har genererar vi ju själva. Alltså, när vi hämtar in data genom inspelning som jag nyss nämnde så är det ju mycket vi skapar själva också, men inte alltid. Vi är med väldigt mycket i inspelningen av flertalet scener och har *motion capture* data och allt möjligt. Bara en inspelning som vi var på för några månader sen hade efter tre veckor runt 500 terabyte data, och det är inte ens en full featurefilm direkt. Men sen genererar vi ju väldigt mycket data. Vi bygger många modeller av saker, exempelvis träd, djur, krukor, hus, ja, *you name it* - lite allt möjligt. Vi bygger muskelvävnader. Vi kör fysiksimuleringar på hur muskler ska röra sig naturligt på exempelvis djur och människor. Även för kläder eller andra material, att allt ska se naturligt ut oavsett vilket material det gäller. Sen kör man även explosion-simuleringar i olika scener; spränger saker eller öppnar stora bränder vilket involverar rök och annat. Allt detta genererar ju otroliga mängder data för varje *frame* (bildruta). Och sen gör vi ju renderingar, alltså bilder av denna datan från digitala kameror. Detta genererar ju otroligt mycket data det med. Så det är väldigt mycket in-house genererat. Och detta lagras ju på centraliserade databaser över hela världen. Vi har större kontor på flertalet platser globalt där alla har lagringsutrymme centralt. All den här datan synkroniseras då vid olika tidpunkter om dygnet, så att alla kontor använder samma data. Och där

har vi ju då ett visst lagringsutrymme, som ska vara till för massvis med petabyte av data såklart… men vi rensar hela tiden och lägger data på långtidslagring också. Man kan säga att vi lägger datan på tape, vilket är en väldigt långsam process men där det ryms väldigt mycket data, vilket då är för backupändamål.

Intervjuare (CD): **Jag flikar in lite med en fråga. Har ni då någon redundans på er data i era datacenter? Ifall någon databas skulle krascha tänker jag så inget går förlorat.**

Respondent: Ja, det finns det absolut. Vi har massvis med backups. Det har vi på allt i alla olika steg.

Intervjuare (CD): **Ja, det låter smart! Men okej, jag tänkte fråga om ni genererar all data själv eller om ni hämtar från tredje part, men det verkar du redan ha svarat på.**

Respondent: Både och, Ja, precis. Allt möjligt. Sen kan vi ju skicka iväg modeller som vi har gjort, och så kan ett annat bolag rendera bilder av detta som vi sedan får tillbaka för att sammansättas. Så det är både in- och utflöde kan man säga. För ibland är det liksom en scen som vi absolut skulle klara av att göra, men det är inte så viktigt att den är "bra", eller att vi kan spara massvis med pengar på att skicka iväg den till andra mindre bolag eftersom de är billigare. Så då har vi sagt till klienten att "det här kommer vi leverera på si och sån budget", där vi sedan skickar till mindre bolag som får oss att spara pengar. Typ som att jag skulle betala någon för att jobba åt mig - fast billigare.

Intervjuare (CD): **Ja, okej. Så nu har du beskrivit extrahering och hur ni främst lagrar datan. Vad är nästa steg?**

Respondent: Juste. Precis. Ja, alltså, vi ligger ju egentligen ganska långt efter, vilket man kanske inte kan tro, just när det kommer till AI (Artificiell Intelligens) eller Big Data, egentligen. Vi har däremot en grej. Vi har ju ett system som tar in statistik om, typ, hur mycket lagring som används vid given tid, och om hur mycket vår *render farm* används till exempel, här finns det massvis med statistik. Till exempel statistik på hur lång tid en rendering tar baserat på massa olika parametrar, och detta används för ett system som listar ut hur det ska schemalägga dessa olika uppgifter på vår render farm. Det är en applikation som vi använder. En annan applikation som jag specifikt har jobbat på är inom brusreducering. Här har vi en massvis med renderingar, alltså färdigställda bilder kan man säga, som tar massor av tid att rendera och blir mindre och mindre brusiga för varje gång renderingen kör, alltså mer detaljrik kan man säga. Om man då renderar jättelänge så blir bilderna väldigt cleana, men det tar väldigt, väldigt lång tid i slutet ju mer detaljerade bilderna blir, och skillnaden blir inte överdrivet stor. Så där har vi ett AI-system som tittar på en brusig version av bilden och själv listar ut hur det ska se ut i slutet av en färdigställd bild, vilket sparar tid. Det är alltså en

maskininlärningsprocess där man har en brusig bild och ett facit, så ska vår applikation lista ut från den brusiga bilden, genom flera upprepningar då, hur bilden snabbast blir fullständig. Träningsdata kallas det i maskininlärnings-världen, som man använder för att göra applikationerna bättre och snabbare. Där uppstår ju problem också typ, hur mycket data ska vi spara? Vi vill ju inte spara renderingar som vi inte använder. Hur skulle det se ut om vi sparar all data för varje dag i all oändlighet? Det går ju liksom inte för det tar för mycket tid och plats. Här är en annan aspekt i, ja, när vi har tagit bort en show eller när en film är klar, då skickas ju allting upp till backup-lagring, så då kan ju inte jag använda det materialet eftersom det ligger på tape. Jag kan liksom inte använda den datan då för att exempelvis träna ett neuralt nätverk, för det tar flera dagar att läsa datan, så då måste vi, liksom, man måste ha det man ska använda på snabb-lagring, men då är det ju dyrare eftersom man inte har lika mycket plats och den måste vara snabbt tillgänglig. Det säger ju sig självt…

Intervjuare (CD): **Juste, precis. Så du skulle ändå säga att ni absolut hanterar Big Data? I alla dessa olika applikationer och processer som du nämnt.**

Respondent: Ja, absolut. Alltså i allra högsta grad!

Intervjuare (CD): **Ja, okej! Då är vi väl i princip klara med bakgrunden. Jag tänkte fråga lite mer specifikt om processerna inom er datahantering men det tycker jag du täckte tidigare…**

[12:49] **IT-verktyg**

Intervjuare (CD): **Nu går vi in på början av vår kärn-del, och det hanterar själva IT-verktygen, exempelvis Hadoop och sådana mjukvaror som ofta används inom hantering av Big Data.**

Respondent: Ja, jag förstår, men vi använder nog mest in-house tools tror jag, jag vet dock inte riktigt hur mycket jag får berätta om de, även om jag hålls anonym haha.

Intervjuare (CD): **Nej, absolut, du får göra ditt bästa! Första frågan är i alla fall; berätta för oss vilka IT-verktyg ni använder inom era datahanteringsprocesser. Och du nämnde precis att det är mest in-house skräddarsydda verktyg?**

Respondent: Ja, precis, det är mest in-house skräddarsydda verktyg för det mesta.

Intervjuare (CD): **Precis. Och varför dessa in-house skräddarsydda verktygen före andra standardiserade verktyg?**

Respondent: Jo, det har nog mycket med legacy anledningar att göra. Att vi har ett, ja eller, det är ju ett väldigt gammalt företag så det är inte bara att slänga ut det man har och byta ut det. Vi bygger nya system hela tiden för de andra

vi jobbar med kanske inte har det vi behöver, eller så har de system som fångar data på ett felaktigt eller annorlunda sätt från oss. De kan vara baserade på andra, kopplade system som även de är skräddarsydda långt bak i tiden osv. Det är mycket sådant. Det är ett stort kluster som påbörjades liksom 30 år sedan, haha, så det är väldigt komplext.

Intervjuare (CD): **Nej, okej, men det känns ändå rimligt i och med åldern på företaget.**

Respondent: Ja alltså, jag är rätt säker på att vi inte använder Hadoop eller sådant liksom, vi har ju inte Big Data riktigt på det sättet, typ Scala Hadoop lösningar och så, eftersom vi inte tar in flera miljarder Likes i sekunden som typ Facebook. Utan vi tar ju in typ ganska medelmåttiga mängder, i Big Data-termer, om exempelvis hur våra servrar och vår render farm används. Det är typ på den nivån.

Intervjuare (CD): **Okej, intressant. Då ska vi se. Okej, inom vilka steg, eller faser, av datahanteringen används era verktyg?**

Respondent: Det är nog inom alla olika steg, eller vad man nu kan säga. Jag tror att vi har ett stort system som sträcker sig liksom över hela vår pipeline. Det är ju inte så många steg egentligen. Vi mäter typ… Ja, alltså det jag sitter med mest så använder vi flera olika Python script som vi har utvecklat med en research-avdelning hos vårt moderbolag för att hantera datan. Så det är ju inte supermycket datahantering i sig, men alltså, där med hanteringen, det är ju egentligen bara flera customfiler i ett stort nätverk där vi skickar filer mellan servrar med hjälp av dessa olika script. Liksom, käll-script.

Intervjuare (CD): **Juste, okej. Så det är alltså dessa olika script som är era verktyg, så att säga?**

Respondent: Alltså allting är ju *custom made* så att säga. Allt är skräddarsytt för våra processer.

Intervjuare (CD): **Okej, och hur ser det ut i företag av er skala generellt?**

Respondent: Jag tror att det liknar sig väldigt mycket på alla möjliga företag i vår skala, det beror på lite, om ni intervjuar någon inom finansbranschen eller inom sociala medier eller så som sparar mycket persondata eller så, så kanske de har mer standardiserade lösningar som Hadoop osv.

Intervjuare (CD): **Jo precis, man kan ju tänka sig hur mycket sociala medieföretag har som resurser inom detta.**

Respondent: Haha, ja de har säkert flera avdelningar delegerade till att enbart analysera varför vissa smileys används eller inte…

Intervjuare (CD): **Haha, ja exakt. Men okej, nu kommer en rolig fråga här, skulle du kunna rangordna till vilken grad (1–5) era verktyg ger er det stöd ni behöver för era processer?**

Respondent: Ja, alltså det är ju lite svårt i och med att vi inte har någon utnämnd lista på våra verktyg eftersom majoriteten av de "verktyg" vi använder är script. Men vi ska ha ett system, jag har bara glömt nu vad det heter, men det ska ta in statistik om olika processer som vi kör. Men det systemet är ju också custom byggt av någon snubbe liksom.

Intervjuare (CD): **Mm, okej jag förstår, men kan vi istället ändra om frågan till att se inom vilka delar, eller faser av datahanteringen, som ni anser fungerar bäst eller mindre bra?**

Respondent: Jo men alltså inhämtning eller generering fungerar ju hur fint som helst, så det är väl en 5a då i rangordningen, haha. Det är finemang. Det är liksom det vi gör mest. Sen är problemet att, vi har ju väldigt mycket potential i vårt företag för att automatisera saker - som vi inte gör fastän vi borde. Vi har typ, alltså, vi kan ta in en massvis med data sen så ska det färgsättas på något vis. Eller man kan ta ett annat exempel, vi har ju vissa avdelningar som gör specifika saker med de renderingarna, alltså bilderna, som vi har färdigställt, som kan omfatta att man klipper ut olika delar i bakgrunden som ska bytas ut med något i en scen. Där sitter de och ritar konturen av vissa objekt för att exempelvis byta ut bakgrunden, men det kan man ju göra automatiskt med AI, och där har vi massvis med data, men vi har inte gjort någonting med denna datan. Så det sitter jag nu med och försöker få ihop. Så just användningen av datan är askasst, tycker jag, potentiellt. Jämfört med vad vi skulle kunna ha det ur ett maskininlärningsperspektiv så är vi superdåliga tycker jag, och det är ju för att vi inte har folk som kan.

Intervjuare (CD): **Intressant! Ja, juste, jag tänkte precis flika in med vad anledningen för att er användning av data är "dålig" skulle vara.**

Respondent: Jo precis, och det är ju för vi har anställda som inte har rätt kunskap inom de områdena så det blir värt att utföra arbetet.

Intervjuare (CD): **Ja, okej, men vad bra! Då är vi klara med det. Då kommer själva kärn-delen nu då, och kommer belysa begränsningar av dessa mjukvaror som vi pratat om.**

Respondent: Japp, absolut.

[19:23] **IT-verktyg**

Intervjuare (CD): **För att vi ska kunna höja reliabiliteten lite så vill jag ändå börja fråga dig om hur du skulle definiera en "begränsning" inom en mjukvara?**

Respondent: Alltså i den mest allmänna benämningen?

Intervjuare (CD): **Ja, men precis.**

Respondent: Ja, alltså, en begränsning är ju att det finns något som ett program inte kan göra. Alltså, att, jag inte kan göra den här saken med det här verktyget, eller att det är så jobbigt att jag istället väljer ett annat verktyg. Då har det verktyget den begränsningen, givet vilket fall det nu än är...

Intervjuare (CD): **Ja, exakt. Okej. Och nu, beskriv gärna eventuella begränsningar inom era respektive IT-verktyg? Så det får ju bli inom era script eller applikationer i sådana fall. Begränsningar inom det ni använder inom er pipeline för datahantering.**

Respondent: Ja, eller alltså, okej, en grej är väl att eftersom vi har så mycket skräddarsydda applikationer och verktyg så kan det bli problematiskt om det kommer nya människor och jobbar. De måste då alltså lära sig allting, man kan ha jobbat på Facebook till exempel med det mesta standardiserade ramverket som även används på Google, och så kan personen inte börja där utan kommer till oss istället. Då kan inte de komma med frågan "Har ni något som heter x? Vi använde y på min senaste arbetsplats". Inga har liksom sett våra verktyg eller processer om man inte jobbat på vårt företag. Så det är ju en begränsning, dock en mer organisatorisk begränsning om det nu är spännande… Sen är det ju också, ja, alltså det blir nästan samma grej, men om vi har en som är ansvarig för ett visst system skulle sluta så står vi där helt *clueless* eftersom vi lagt alla våra ägg i samma korg. Det är ju också en begränsning.

Intervjuare (CD): **Ja okej, dessa är ju hyfsat organisatoriska begränsningar som du nämnde. Vet du några specifika tekniska begränsningar?**

Respondent: Ja, jag kan ju mest prata om vad jag sitter med och vilka begränsningar verktygen har inom mina arbetsområden … Jag försöker tänka vad det är som är så dåligt egentligen, haha. Det skulle väl i sådana fall vara att vår kod som vi använder är riktigt usel. Den är inte produktionsvänlig för fem öre eftersom det varit ett forskarteam som suttit och skrivit den, inga direkt professionella utvecklare som har kodande som deras huvudsyssla. Den är liksom inte modulär. Den fungerar och är bra, men den är inte utvecklad för att kunna hantera specialfall på ett optimalt sätt. Det är ofta sådana här problem med hembyggen, de är inte så fullständiga, inte så polerade. Det är en begränsning inom vissa av våra applikationer, jag kan inte säga på rak arm några specifika applikationer eller script, men det är min uppfattning.

Intervjuare (CD): **Okej, så du menar att applikationerna ofta inte är fullt optimerade till alla möjliga fall?**

Respondent: De är inte så polerade helt enkelt, de är inte så förberedda för specialfall. Jämför i allmänhet med att använda ett stort standardiserat verktyg, då kan ju exempelvis ett tillkortakommande med ett stort standardiserat verktyg vara att de inte kan hantera olika specialfall. Men om man har ett hemmabygge så kanske ens gränssnitt är klankigt eller att verktyget är långsamt att använda, kanske kan vara buggigt eller bara massa kluster av skräddarsydda lösningar… Det är ju mest script det handlar om, så det är väl i sådana fall icke optimerade script som skulle kunna vara en begränsning. Att varje script kan utvecklas till att fungera snabbare eller så, men de rör ju egentligen bara filer från en plats till en annan så det är inte så mycket som kan gå fel där. Men det kan ju tillkomma specialfall som vi kanske inte tänkt på, så det skulle väl vara en begränsning - oförberedda för specialfall eller ej optimerade script i grund och botten.

Intervjuare (CD): **Ja okej! Vad blir då effekten av den här begränsningen på er organisation?**

Respondent: Ja, men effekten kan då vara… Det blir svårt att liksom, det är svårt att utbilda nytt folk, och om folk drar så är det inte så bra för då kanske, som jag sa tidigare, de var de enda med kunskaper inom vissa verktyg. Det är ju mer operativa problem eller risker. Att folk, typ om jag skulle dra från mitt företag så kan inte en enda något om något jag suttit med. Då skulle de var helt körda. Och det är ju inte så bra ur ett företagsperspektiv att ha en "*single point of risk*". Det är ju dock inte mjukvarans fel, men ger då en operativ eller organisatorisk effekt.

Intervjuare (CD): **Precis. Ja men vad bra. Det här har du dock touchat på lite innan, men i och med dessa begränsningar som verktygen har, vad hindrar då er organisation från att byta verktyg?**

Respondent: Kostnad! Kostnad kontra vad vi får ut av bytet. Om vi skulle vilja byta till ett nytt verktyg eller system så får vi ju short-term ingen vinst av bytet. Det kan vara så att nya verktyget fungerar likadant och kan göra liknande saker, men vi måste anställa 4 utvecklare och göra det i två år för att få det att bli funktionellt. Då kan man fråga sig "Vad har vi fått ut av bytet?". Då skulle svaret vara "ingenting" med motivation till "om den enstaka personen skulle dra så skulle vi vara kvar med ingen kunskap alls, med det här bytet minimerar vi risken". Och så här är det ju med allt. Alltså, ja, det är en kostnadsfråga kontra när vi ser återbäring.

Intervjuare (CD): **Hehe, precis, okej. Det finns också alltså lite risk att… Ja, det du nämnde tidigare ...**

Respondent: Liksom, om du skulle investera allt du äger i någonting med kännedomen av att du blir mångmiljonär om 6000 år av den investeringen. Då är det liksom inte så stor anledning till att göra det, haha.

Intervjuare (CD): **Haha, nej det skulle man ju helst vilja avstå ifrån. Men som sagt, det du nämnde tidigare, att ni använt mycket legacy system som**

varit med länge, om ni byter ett sådant, vad skulle effekten av det vara?

Respondent: Jo men absolut, precis. Alltså, allting är ju som en väldigt stor ihopvävd sak, så ifall vi skulle byta ut något så fallerar hela nätverket, haha. Så om vi skulle vilja byta ut ett legacy system så skulle det säkert förstöra kompatibiliteten av flera andra system. Det här är ju en begränsning det med. Att de olika verktygen och systemen vi använder inte är så dynamiskt kompatibla med andra, nyare verktyg.

Intervjuare (CD): **Okej, jättebra. Det gäller alltså alla era verktyg och system liksom?**

Respondent: Ja, precis, det kan vara så att vi vill uppgradera ett verktyg som vi använt för att göra något snabbare, men den uppdateringen går inte att göra för då måste vi uppdatera ALLA andra system eller verktyg som jobbar i symbios eller måste vara kompatibla med det verktyget. I alla olika divisioner av de företagen vi jobbar med, liksom, och det är inget de vill betala för bara för att göra ett verktyg lite snabbare… Och det är liksom, det hade varit mycket bättre, eftersom allt vi gör skulle utvecklas snabbare och bli mycket snyggare, typ bilder osv, men, då blir det kaos. Vi kan göra bytet lokalt, men då kan vi inte använde de globalt sen för det kommer bryta kompatibilitet mellan andra verktyg, system eller så. Man har inte uppdaterat hela pipelinen alltså.

Intervjuare (CD): **Där ser man, okej. Så för att summera då så kan man säga att kostnad och kompatibilitet är anledningar till att ni inte byter ut "begränsade" verktyg?**

Respondent: Ja, precis, det skulle jag nog säga. Framförallt kostnad, det ligger ju i grund och botten till allt. Hade vi vetat, om vi skulle gjort bytet eller uppdateringen globalt, att vi skulle börja tjäna tre gånger så mycket pengar efter den förändringen - då skulle vi ha gjort det. Men nu istället är det, ja, bilderna kommer bli 5–10% finare med uppdateringen av x verktyg… "Inte riktigt värt det", säger högsta hönsen då, haha.

Intervjuare (CD): **Juste, tyvärr svårt att veta innan, hehe. Okej. Hur tror du då att ett plötsligt behov av att hantera en ökad mängd data skulle påverka användandet av era respektive IT-verktyg?**

Respondent: Jag tror nog vi hade behövt… Alltså, om det kom ett behov av att vi skulle behöva hantera mycket, mycket mer data, vilket inte är så sannolikt att det skulle hända, men om något sådant skulle ske så tror jag nog att det hade varit en bra idé att försöka standardisera verktygen, typ använda Hadoop eller de verktygen som stöds av Hadoop. Vi har ju större mängder data när det gäller scener och sådant, men det kommer aldrig komma upp i nivåer som Facebook hanterar exempelvis. Men om vi nu helt plötsligt skulle nå upp i sådana mängder så skulle vi nog behöva tänka om helt tror jag. Och där måste man ta med aspekten igen att man måste se någon vinst på det, vilket i början, återigen, kommer innebära en större kostnad.

Intervjuare (CD): **Om vi då säger att ett krav kom på att ni skulle producera mycket snyggare bilder, hur skulle era respektive verktyg kunna hantera det?**

Respondent: Det skulle inte vara någon skillnad alls. Inte alls. Det beror på att vi… Alltså, formatet på vår data som vi använder förändras inte med volymen. Alltså, det ökar i volym, absolut, men jag tror inte det påverkar våra verktyg - utan bara hårdvaran, alltså hur mycket vi behöver av hårdvaran. Tänk dig att du har en bild med 10 pixlar, om man ska göra den bilden mer detaljrik så blir ju pixlarna bara mindre och fler, det är inget annat som krävs egentligen, så länge filerna vi hanterar genom vår pipeline bibehåller samma struktur och format så kommer det ju inte påverka verktygen, om inte, som tidigare sagt, volymerna blir för stora att hantera inom den tidsgränsen man sätter upp för vardera saken… Ja, du fattar.

Intervjuare (CD): **Yes, precis. Men då har vi fått klart begränsningar-delen, så då kommer vi som slutsats att blicka framåt lite in i framtiden.**

[33:32] Framtiden

Intervjuare (CD): **Yes, framtiden då. För dig och företaget, vilka är då, skulle du säga, era kortsiktiga utmaningar inom er Big Data-hantering?**

Respondent: Ja, alltså, jag vill ju ha ett mer automatiserat sätt att samla in träningsdata från vår renderingsprocess. Och ett mer väldefinierat. Just nu är helt kaos. Min utmaning är att vi får in mer data-tänk i hela vår pipeline, att typ att öka antalet insamlingsprocesser, alltså fler sätt att samla in datan på. Vi har inte så mycket, tyvärr, det är det som är grejen. Vi får ju in data, men det kanske inte riktigt är det jag menar, men det är mer att organisera och *labela*, alltså kategorisera och strukturera datan. Vi skulle, alltså, vi har massa processer. Om du ser hela Företag A som en stor process; vi har asmycket processer, men vi har inte så mycket Data Mining på gång. Det är inte så mycket som strukturerar upp.

Intervjuare (CD): **Du menar alltså att ni har väldigt mycket ostrukturerade data liggandes?**

Respondent: Ja, eller alltså det kommer in massvis med data, sen används det inte och sen sätts det på lagring. Vilket är en *opportunity cost*, alltså en potential att tjäna pengar eller få förbättring inom ett område, men som inte utnyttjas.

Intervjuare (CD): **Jo, precis, men okej. Då är vi framme vid slutet! Är det något vi har missat att ta upp tycker du?**

Respondent: Nej, tycker ni löste det där galant! Hehe.

[37:05] **Avslutning**

Intervjuare (CD): **Då signalerar jag nu att vi stänger av transkriptionen.**

Appendix 3: Intervju 2

Intervjuare: Christian Dahlberg (CD), Rikard Funck (RF)
Respondent: Mats Dahl
Organisation: KPMG
Datum, tid och plats: 16 april, 16.00, Lund (Sverige)
Intervjutyp: Digital intervju över Microsoft Teams
Tid: 48.55

[00:00]	**Introduktion**

Intervjuare (RF): **Okej, sådär! Anonym eller inte anonym?**

Respondent: Nej men jag behöver inte vara anonym.

Intervjuare (RF): **Alright.**

Intervjuare (CD): **Inte heller din organisation? Kan vi nämna det också?**

Respondent: Det kan ni göra, absolut, så ska jag försöka vara så företagsmässig som möjligt.

[00:45]	**Bakgrund och Big Data management**

Intervjuare (CD): **Okej då, då kan vi börja med lite bakgrund kring dig, vem du är, vilken roll du har på företag och hur länge du jobbat där?**

Respondent: Mm, precis ja. Jag är ansvarig för någonting som kallas för Lighthouse inom KPMG, och Lighthouse är ett kompetenscenter för datadrivna teknologier som det heter, om man översätter det från engelska till svenska, och det innebär att vår roll är att hjälpa våra revisorer, våra skatterådgivare och våra rådgivare, managementkonsulter, riskanalytiker och andra att helt enkelt digitalisera deras nuvarande och framtida arbetsprocesser. Jag skulle vilja påstå att 80-90%, 80% i alla fall av det vi gör, oavsett om det är revision eller om det är skatterådgivning eller om det är rådgivning, på något sätt baseras på data och mycket av den datan vi arbetar med idag och den databearbetning som görs är via Excel, vilket är något som man gjorde redan när jag började inom IT-industrin på 1990-talet så att det vi egentligen är inne i är en ganska stor transformation som alla andra bolag och tjänsteföretag dvs. använda moderna verktyg och Big Data. Istället för att göra analyser på stickprov, kunna hantera stora mängder data istället och automatisera processerna, minska antalet timmar men öka innehållet i analysen och förmågan att göra bättre analyser så att det blir ett högre och större kundvärde på många sätt och på det sättet också kunna växa vår verksamhet totalt sett med hjälp av att vi skapar ett större kundvärde och andra typer av erbjudanden. Vi som alla andra bolag eftersträvar, eftersom vi jobbar med enorma datamängder på många sätt, att bli ett datadrivet bolag att

använda data som ett sätt att affärsutveckla vår affär och som sagt att ge våra kunder än mer värde baserat på den datan och de insikter vi har men också alla våra experter som med sin intelligens och erfarenheter ytterligare adderar värde på data insikterna. Lighthouse-verksamheten är inget som är svenskt i sig, vi är runt ett 30-tal personer och på väg att bli 40, i norden så finns det också Lighthouse-verksamheter och totalt sett så är vi nästan 400 dataspecialister i norden inom KPMG och i Europa är vi 7000 och i världen är vi 14000 så hela det här Lighthouse konceptet är ett strategiskt koncept för att lyfta hela KPMG globalt i 154 länder, 220 000 medarbetare, och syftet är att vi har ett gemensamt sätt att använda data och utveckla gemensamma tjänster så att våra kollegor i Sydafrika rent teoretiskt kan utveckla en bra och smart tjänst som vi i Sverige använder några veckor senare eftersom vi använder samma plattformar, samma typ av kompetenser och förmågor och på det sättet att vi får en global hävstång i den förändringsresa vi är inne i vilket gör att vi ytterligare i varje medlemsland kan bli ännu mer konkurrenskraftiga. Vi i Sverige är lite sena i utvecklingen i jämförelse med våra nordiska kollegor, och världen övrigt men vi satsar rejält nu på detta utifrån ett svenskt perspektiv. Jag har jobbat inom IT-industrin sedan 27 år tillbaka och jobbat mycket med gränslinjen mellan verksamhet och IT, jag är ekonom, inte utvecklare i grund och botten och vet vad datan ska användas till utifrån ett affärsmässigt perspektiv, men dem som vi rekryterar till Lighthouse är Data Scientist, Data Engineers, det är utvecklare både front-end och back-end, det är statistiker och andra typer av kompetenser som då kan både använda, omvandla och skapa värde för datan tillsammans med våra domän-experter inom revisionsskatt och rådgivning. Vi har alltså en global strategi, vi bygger verksamheten gemensamt och den plattform som vi använder för att hantera Big Data det är Microsoft Azure och KPMG globalt gick ju ut med en pressrelease i början av december förra året där man meddelade att vi kommer att investera 5 miljarder dollar de kommande fem åren, alltså 50 miljarder svenska kronor, 10 miljarder per år i att bygga digitala tjänster baserade på Microsofts plattform, själva strategin är rätt tydlig med Azure men däremot så har vi ju självklart globala samarbeten även med Amazon, Google, IBM, Oracle, alla dem stora IT-leverantörerna och även mindre inom alla tänkbara områden, men för oss själva vad gäller att förbättra oss och bli mer digitala är den gemensamma plattformen Microsoft Azure och då är vi tillbaka till mitt exempel med om kollegorna i Sydafrika utvecklar något så gör dem det på Azure-plattformen, i och med att det är molnet har vi lätt att komma åt det direkt och bara använda det och kan lagra datat utifrån ett svenskt perspektiv till exempel. Det vi har gjort nu också utifrån det här perspektivet att vi kan hantera stora datamängder är att vi bland annat byggt en stor och ny revisionsplattform som just nu heter NASA-plattformen, ett nordiskt samarbete, där vi i Sverige nu sakta men säkert håller på att implementera det här hos oss och vi kommer att revidera alla våra 25 000 små/medelstora revisionskunder i den här molnbaserade plattformen som gör att vi kan automatisera delar av revisionsprocessen och vi kan, som sagt, skapa helt nya typer av insikter, vi kan revidera hela populationen istället för att göra stickprov, vi kan skapa insikter som kanske inte kunder eller vi haft

tidigare och vi kan samordna och jämföra marknader eller industrier och företag på ett helt annat sätt än tidigare. Det här är ganska nytt så vi håller på att implementera det och där vi också agera med kunden utifrån ett digitalt perspektiv dvs att datat kommer till oss så att säga, både genom att kunden loggar in på våra nya hemsidor: MyKPMG och annat, där kunden i inloggat läge med BankID kan ladda upp sina filer så att vi har kontroll på vem som laddar upp vad, men där vi också nu tittar på hur vi kan integrera oss med våra kunder via standardiserade API:er så att vi kan få ett kontinuerligt dataflöde och att vi rent praktiskt så småningom istället för att göra revisionen en gång per halvår rent teoretiskt skulle kunna göra revisionen varje dag eller en gång i veckan eftersom vi har flödet av data in till oss. Det är en ganska stor förändring och vi har precis i början av det, strategin är på plats, vi är på plats, men vi har en förändringsresa framför oss de kommande åren som jag tror kommer vara väldigt spännande för alla parter. Det sista jag vill säga i den här introduktionen är att vi inte bara gör det här för att vi ska utan för att våra kunder kräver det av oss, om vi inte gör dessa sakerna kommer vi inte ha våra kunder kvar, dem förväntar sig att vi ska kunna samarbeta digitalt, både i relationen men även rent datamässigt, dem förväntar sig i och med att vi har stora datamängder att vi kan ge dem andra insikter än bara en revisionsrapport eller revisionsberättelse t.ex. men det gäller även för våra skattekunder och rådgivningskunder i övrigt, så det är väl i stort vad vi håller på med och vår strategi, och det jag säger nu är det vi säger till allt och alla för att beskriva att vi är på en väg framåt där vi tror vi kan vara disruptiva på denna marknaden vi befinner oss på.

Intervjuare (CD): **Spännande! Jag tänkte att innan vi kommer in på Big Data mer tekniskt, hur skulle du definiera just Big Data lite kortfattat?**

Respondent: Det är en bra fråga! Jag skulle definiera det som att Big Data är lika med att man kan använda obegränsad mängd data för en obegränsad mängd olika typer av tjänster och insikter som kan driva och förändra både affärsmodeller och verksamheter, tillbaka till mina exempel här kring revisionen, när vi jobbar med Excel-ark så har vi en begränsning på en miljon rader men när vi gjorde revision på en av de större butikskedjorna här, när vi hjälpte till från Lighthouse sida, så hanterade vi datamängder motsvarande runt 20-25 miljoner transaktionsrader och där vi också ska jämföra t.ex., som inte revisorerna kunder göra tidigare, en trevägs-matchning mellan kassakvitton, bankinsättning och lageruttag där det handlar om runt 25 miljoner transaktioner per styck och där vi kan på sekunden säga att dem här transaktionerna finns det ingen trevägs-matchning för, vi kan ju gå ner på transaktionsnivå och säga att Mats Dahl var i en viss butik och köpte en chokladkaka, han betalade med kortet, här var det bankinsättningar och lager uttaget, alla dem transaktionerna på den nivån kan vi verifiera och så kan vi också titta på där det var ett lageruttag men inget kassakvitto och vice versa. Det var helt omöjligt tidigare innan vi använde kapaciteten på Azure-plattformen där det är obegränsat lagerutrymme osv och där analysen görs blixtsnabbt, för mig är alltså Big Data den obegränsade möjligheten att använda gigantiska datamängder för att skapa insikter och värde framåt.

Intervjuare (CD): **Skulle du kunna förklara en översikt på er allmänna process för hantering av Big Data? Vi syftar då på de olika stadierna i processen, samla in data, lagra data, processa data, analysera och visualiserar data.**

Respondent: Det är lite som du säger. För vår del, enkelt uttryckt, i och med at KPMG globalt har valt att använda Microsoft Azure så använder vi alla dem mikrotjänster som finns inom ramen för Azure dvs. Saker som *Data Factory* och *Data Bricks* för att både extrahera data och lagra den på ett vettigt sätt. För oss är det extremt viktigt att den datan som vi tar in är kunddata och inte vad som helst, att ingen annan än dem som ska komma åt datan gör det, exempelvis våra revisorer eller våra skatterådgivare, vi har en väldigt strikt hantering av *active directory* mm. För att säkerställa att ingen kommer åt datan, att vi lagrar den på ett sätt som gör att det är säkert dvs. Hela vår Microsoft-uppsättning bygger på att vi är helt kompatibel till Microsofts beskrivning av hur man sätter upp en sån här miljö, som ni kanske vet så är Microsofts hela miljö en tjänst till många så det finns inget undantag varken för oss eller, som exempel, för amerikanska försvarsmakten att ändra i tjänsterna som Microsoft har, det är samma för ex. Google och Amazon. Vi har sett till så att vi är helt kompatibla dvs. Det finns massa dashboards som beskriver om vi satt upp våra olika typer av tjänster på ett sätt som är enligt Microsofts riktlinjer, det finns alltså en mängd indikatorer som är gröna, gula eller röda som visar huruvida vi satt upp det rätt eller inte och dessa indikatorer i sig är kopplade till Microsofts certifikat eller ISO. Vilket då innebär att vi helt arbetar efter Microsofts standard och vi blir då också per automatik både SOCKS och följer också ISO. Det vi ju också gör för att säkerställa att våra miljöer är korrekta och att vi begränsar alla risker man kan tänka sig är att vi i alla lösningar vi bygger, t ex revisionsplattformen, att vi har en tredjepartsgranskare som är expert på Microsofts plattform som kollar att dem uppsättningar vi har, där vi byggt själva lösningen, att dem också följer standarder och best practices så att vi även där har ett intyg på att allt vi har gjort har följt alla dem regler och riktlinjer man ska hänsyn till. Om vi följer allting är det Microsoft som har ansvaret för att säkerställa att säkerhet och eventuella intrång inte sker. Vi är väldigt måna om att vi har rätt kontakt med våra kunder och får in datan på rätt sätt i en bra miljö och ser till att vi ger tillgång till den här datan på olika sätt till de som ska ha den. Vi använder just nu bara Microsofts olika tjänster, visualiseringen gör vi t.ex. med Power BI och det hänger ihop med att vi vill ha en helhet i allt vi gör och inte håller på att mixtrar eller gör på olika sätt vid olika typer av lösningar utan att vi har ett gemensamt tillvägagångssätt. Det vi också har gjort nu är att vi är tydliga med att vår IT-avdelning också är involverade i allt detta, vi har fördelat ansvaret från IT via oss på Lighthouse och till verksamhetsområdena så att alla är involverade i hur datan lagras och används. När vi använt datan enligt de överenskommelser vi har med våra kunder stryker och kastar vi denna efter vi använt den vilket är viktigt i vår bransch men också att vi vill, i så stor utsträckning som möjligt, behålla kunddatan och anonymisera denna för att kunna göra andra typer av analyser och insikter på stora

mängder kunddata, vi vill ju kunna lagra hur mycket data som helst för att själva kunna hitta synergier eller trender osv.

[18:01] IT-verktyg

Intervjuare (CD): **Utmärkt! Då går vi in lite mer på just IT-verktyg, du nämnde Azure ganska mycket och just Microsofts produkter men skulle du möjligen kunna redogöra för vilka IT-verktyg ni använder för de olika stegen i er datahanteringsprocess?**

Respondent: För det första, anledningen till att KPMG har valt Microsoft Azure som våran strategiska molnplattform har att göra med att Microsoft tillhandahåller infrastrukturen i molnet och alla de tjänster som jag snart ska räkna upp, men Microsoft äger inte kunddatan, om vi tar, exempelvis, Excel, innan implementerade man Excel på sin lokala dator och hade datorn både som infrastruktur, lagringsplats och applikationsserver och så körde man Excel på den, så Microsoft har nu sedan 4-5 år tillbaka ändrat hela sin affärsmodell och säljer Excel som en digital tjänst och äger allting, ansvarar för säkerhet osv. Men kommer inte åt det du faktiskt lägger i den enskilda cellen utan den datan äger vi som kund och det är extremt viktigt för oss som revisionsbolag, som hanterar stora mängder kunddata, att ingen annan kommer åt det, ingen kan tjäna pengar på det osv. Vilket är fallet till viss del med dem andra stora molnleverantörerna att dem har en annan affärsmodell än Microsoft har och det kan inte KPMG globalt som revisionsbolag acceptera utan för oss är det viktigt att vi har kontroll på kunddatan och ingen annan kan se den och är det så att t.ex. att någon, typ, amerikanska staten säger till Microsoft: vi vill se all den kunddata som KPMG har så säger Microsoft till dem, vilket de gjort i tre stycken rättsfall: vi kan inte för vi skulle då stjäla den datan från KPMG och även om dem skulle skänka det eller stjäla datan så kan inte någon annan läsa det eftersom det är vi som har krypteringsnyckeln till den, för oss är detta extremt viktigt. En annan anledning till att vi valt Microsoft är att de tjänster som vi använder, applikationerna, redan finns där t.ex, det här med data factory är ett standardiserat sätt att hantera och integrera data, vi har hela lagringsdelarna kring allt ifrån olika typer av SQL-databaser osv. Både för strukturerad data men också för ostrukturerad data, vi har som sagt möjlighet att skapa data-sjöar med Data Bricks, vi har Power BI som visualiseringsverktyg, vi har redan pratat med kunder att vi kan tillhandahålla tjänster kring IOT t.ex. via en IOT-hubb som redan finns tillgänglig, vi jobbar också med Azure bot-services för att bygga botar, vi använder Microsoft AI och Machine Learning funktionalitet för att bygga lösningar kring det och även det som kallas för Natural Language Processing (NLP), dvs hur vi går till väga för att läsa stora datamängder. För vår del, istället för att vi köper alla dessa komponenterna i form av olika typer av applikationer och lägger på någon form av serverstruktur, det skulle kunna vara Azure eller på Amazon, vi skulle fortfarande kunna köra det i molnet, så har vi valt att utgå från att vi använder Microsoft olika tjänster inom ramen för Azure för att bygga våra lösningar t.ex. tidigare nämnd revisionsplattform där vi då också använder saker som SharePoint-online

och andra delar som finns inom ramen för Microsoft. Så istället för att vi har massa olika licenser i ett garnnystan av olika applikationer som alla använder på olika sätt så försöker vi hålla oss till de standardfunktioner som finns inom ramen för Azure, dels för att följa Microsofts utveckling av lösningar och att vi själva ska undvika att ha allt för mycket eget utvecklade produkter utan vi använder digitala tjänster istället. Sen får man inte glömma hela denna cyber-biten som självklart är viktig för oss, vi använder hela den säkerhetsprocedur som finns runt Azure i sig och våra Global Head of Audit säger retoriskt att Microsoft investerar ungefär 1 miljard dollar bara i cybersecurity kring sina molnbaserade lösningar, KPMG lägger då kanske runt 1 miljard dollar i vår nuvarande IT-infrastruktur inklusive hårdvara osv. Vem är då bäst på cybersecurity? Det är självklart Microsoft och det är också därför vi väljer dem. Så det är därför vi jobbar inom ramen för denna miljö, självklart kommer vi att samarbeta med Amazon och Google och gör det redan till viss del i vissa sammanhang men då är det mer att Google har något bättre funktionalitet av något slag eller att våra kunder använder Google och vill använda en sådan lösning i vissa tjänster men för våra egen utveckling och vår egen digitalisering så utgår vi ifrån Microsoft och Microsoft Azure plattformen och de tjänster som finns inom ramen för det.

Intervjuare (RF): **Du har flertalet gånger nämnt att ni använder er av verktyg inom ramarna av Microsoft Azure, hade man kunnat gå in mer specifikt på verktygen ni använder inom Azure-plattformen för att få ett namn på dem? Azure, såsom vi förstår det, innehåller ju en större mängd verktyg.**

Respondent: Jag kan skicka en liten uppsättning av vad vi använder, är nog lättare än att räkna upp alla här.

Intervjuare (RF): **Gärna! Vi vill gärna bara ha ett namn på dem, det gör det lättare för vår undersökning.**

intervjuare (CD): **Hade du kunnat rangordna de verktyg som ni använder nu på en skala 1–5 utifrån hur mycket stöd de ger till den verksamhet ni driver med dem?**

Respondent: Det är en väldigt bra fråga, just nu uppfyller dem i princip alla kraven som vi har men kommer säkert att springa på massa saker där det inte riktigt är tillräckligt bra. Microsofts saker är i vissa fall väldigt långt i framkant andra är hyggligt bra i jämförelse med dem som har spetsfunktionalitet, vi tittar dock mer utifrån perspektivet att det är tillräckligt bra och det är bättre att hålla ihop helheten och använda något som kanske inte är marknadens bästa funktionalitet men det är tillräckligt bra för att passa oss. Det är där vi är just nu utifrån molntransformationen vi går igenom, sen är det en utmaning för oss och det tror jag det är för alla när man går till molnet, att man avhänder sig en viss typ av kontroll till Microsoft i detta fallet vad gäller just säkerhet och vem som hackar och vem som kommer åt data osv. och där behöver vi både för vår egen skull men även för alla revisorer, dem som faktiskt jobbar med kunderna

och kundernas data och garantera kunder att vi kommer hantera deras data korrekt, att vi är säkra på att allt är korrekt och att vi verkligen kan garantera kundernas datas säkerhet, där är vi fullt nöjda med det som finns eftersom vi har valt det utifrån ett globalt perspektiv, men jag tror att för den gemene specialisten inom KPMG vad gäller de olika specialistområden, behöver det också säkerställas att dem litar och tror på det här för att jag tror att alla verksamheter oavsett om man tillverkar bilar eller landsting eller konsultföretag som vi så är det här skiftet från att man haft sin data på antingen egna servrar eller outsourcade servrar där man haft ett till ett förhållande till sin outsourcing leverantör, när man kliver in i molnet så blir man en av många som använder samma tjänst, hur kan den här molntjänsteleverantören garantera att dem inte blandar ihop data mellan olika företag osv. Och det är den stora förtroendetröskeln som vi och alla andra behöver komma över, det tror jag är den största utmaningen just nu. Varför KPMG har valt Microsoft är för att dem inte tjänar pengar på att använda kundernas data eller hur kunderna använder sin data, den affärsmodellen måste finnas kvar för att vi på KPMG ska fortsätta använda deras tjänster, det är också väldigt viktigt att vi har en sådan partner.

[19:23] **Begränsningar**

Intervjuare (RF): **Om det skulle tillkomma ett behov av att hantera en ökad mängd data i framtiden, hur skulle det då påverka de verktyg som ni använder just nu?**

Respondent: Min uppfattning är att dem definitiv skulle göra det och utifrån Microsofts perspektiv är det hela deras affärsmodell. Microsofts globala VD, som heter, Satya Nadella han har jobbat på Microsoft de senaste 20 åren och varit med och bland annat utvecklat Azure, men när han blev VD för 4-5 år sedan så öppnade han upp hela Microsoft och gjorde om hela affärsmodellen från att vara en ganska trygg affärsmodell som var att sälja licenser som kunde vara allt från serverlicenser till Excel-licenser som varje företag fick implementera i sina egna miljöer, nu har man istället ändrat affärsmodell och säljer t.ex. Excel som jag sa tidigare och infrastruktur som en digital tjänst. Om ni startar ett företag idag och bygger en Uber-app kan ni bli globala omgående eftersom allt finns där redan, det är bara att trycka på knapparna och sätta upp tjänsten och börja sälja det. För Microsofts del så är hela deras framtid att kunna hantera gigantiska mängder data och det är samma med Amazon, Google och alla andra, det är därför de bygger datacenter över hela världen, Microsoft har 54 så kallade Azure-regions och man bygger tre i Sverige nu, ett utanför Gävle, ett utanför Sandviken och ett utanför Staffanstorp nere i södra Sverige. Amazon har etablerat tre nya datacenter ett i Eskilstuna, ett i Katrineholm och ett i Västerås. Google håller på och prospekterar i dalaskogarna för att bygga datacenter och Facebook har sedan 5–10 år tillbaka tre datacenter utanför Luleå. Allt det här hänger ihop med liknande datacenter över hela världen och det är därför den här datakraften är så gigantiskt stor och det här kommer byggas ut mer, det kan finnas en negativ sida av detta för att dem konkurrerar ut globala

spelare såsom Evry m.fl. som har outsourcingtjänster idag, traditionellt, som man haft i 30 år, dem kommer försvinna helt, när datacentret utanför Staffanstorp är klart om något år så kan jag nästan garantera att region Skåne, skånska/sydsvenska kommunerna Tetra Pak och andra som har en Microsoft strategi idag i sina outsourcade tjänster, dem kommer lägga allt i det datacentret där istället. Det är en gigantisk förändring i IT-marknaden nu, både att traditionella outsourcing-partners försvinner men även traditionella IT-bolag konsolideras ju upp hela tiden, Acando gick upp i CGI till exempel osv. Därför att marknaden förändras, därför att den här typen av tjänster som jag redogör för här är så lättillgängliga, du behöver inte ha en IT-tekniker utan du kan vara som jag, en ekonom och så kan du börja använda grejerna, eller att som vi: rekryterar data scientist och andra och säga att det är den här plattformen vi ska jobba med thats it vi behöver inte ha några servertekniker och liknande. Alla dem här stora molnleverantörerna bygger alltså för att kunna hantera oändliga mängder data och hitta olika former av datacenter, ett exempel är Microsoft som testar datacenter som ser ut som stora ubåtar som dem sänker ner i havet för att få kylning osv. Vattenfall är nu inblandade i Microsofts etablering av datacenter runt om i Sverige där Microsoft nu har sagt att man 2030 inte ska ge något som helst klimatavtryck och om man kollar ytterligare 10 år framåt så ska man även ha sugit in så mycket av all dem utsläpp man haft under de 40-50 åren som man funnits så att man ska vara helt nollställd och Vattenfall jobbar nu väldigt intensivt med de tre datacenter Microsoft bygger för att just leverera fossilfri el till dem via vattenkraft och vindkraft osv. Det är en ganska intressant total utveckling vad gäller att kunna tillhandahålla data, göra det hållbart och globalt.

Intervjuare (CD): **Intressant! Vi går vidare till att fokusera lite mer definitivt på begränsningar ni upplever med mjukvaran ni implementerat för att hantera Big Data, men först vill vi fråga hur du skulle definiera en "begränsning" inom en mjukvara?**

Respondent: Bra fråga! Jag tror såhär: För det första, en begränsning är ifall en programvara som hanterar stora mängder data är skriven på ett sätt som gör att det går väldigt långsamt att processa data, till exempel om man skriver en Pythonkod som gör att det tar 24 timmar att göra en beräkning som man egentligen kan göra på 24 minuter. En begränsning ligger i förmågan och kunskapen i att bygga applikationer som just kan hantera stora mängder data. En begränsning i det fallet tror jag också handlar om att det finns så mycket gammalt tänk kvar, den här exponentiella utvecklingen som skulle kunna ske sker inte därför att man är lite för gammaldags i sitt sätt att se på saker och ting. En annan begränsning är att: vågar man släppa loss folket, jobba med stora mängder data, vågar man tillsammans med kund jobba tillsammans med stora mängder data eller är till exempel vi rädda för att någon tar sig in via våra API:er och förstör för oss utifrån ett säkerhetsperspektiv. Det kan också vara en begränsning att konflikter kan uppstå i alla olika flöden av data som man har inom organisationen. En annan begränsning i dessa tider är standardisering av data. Om man nu ska kunna använda stora

datamängder och integrera processflöden, Supply Chain osv. så måste datan vara standardiserad, en kund är en kund oavsett vilket system den kommer ifrån, ett annat exempel är hur man anger svenska kronor osv. Det är med andra ord svårt att ena sig om hur saker och ting ska vara och där man kanske behöver ena sig på en global nivå vilket tar tid.

Intervjuare (CD): **För att fortsätta på spåret med begränsningar, har ni några upplevda begränsningar inom de IT-verktyg som ni använder för datahantering inom er organisation?**

Respondent: Inte på det sättet, men det är snarare det jag försökte säga innan, just hur vi optimerar det vi gör. Hur får vi olika individer att vilja och välja att jobba tillsammans. Som exempel, revisionsplattformen, och hur vi definierar data från kund och samtidigt jobbar med en annan tillämpning för en annan del av KPMG och då måste vi titta internt så att vi har en fungerande "common data model". Jag tror att det är lätt för oss att sitta och prata om men det är svårt att få experter och individer som inte jobbar så mycket med IT att förstå att detta är jätteviktigt och att man måste fatta beslut kring det. Vidare att vår ledningsgrupp måste förstå att detta är en väldigt strategisk fråga, att vi kommer överens om hur vi ser på data och hur vi kan integrera data så att vi kan bygga applikationer som snabbt ger ett värde till kund. Också att om vi hjälper kunder på ett ställe så kan vi använda samma data för att göra insikter i en annan del av vår verksamhet. Dock, i och med att vi mer och mer använder oss av Microsoft-baserade tjänster, som vi har inga specifika begränsningar i dessa, men det är just hur vi tillämpar det och hur vi standardiserar vårt sätt att arbeta, det kan bli ganska tråkigt för många för man kanske är van vid att kunna kasta in en applikation här och en open-source där och själv skapa en databas osv. Men i den här digitala världen så kommer det att bli mer och mer struktur och krav och modeller som man ska följa, jag är helt övertygad om att till exempel när svenska myndigheter, kommuner, landsting och regioner upphandlar så småningom att dem kommer ställa vissa krav på att applikationer och tjänster ska vara på ett visst sätt så att dem kan integrera sig med varandra. Ta som exempel, nu i dessa Corona-tider, Socialstyrelsen jobbar hårt för att få ett grepp om var allt material finns inom hälso- och sjukvården vad gäller t.ex. skyddsutrustning, respiratorer osv. Man vill ju ha ett kontrolltorn från socialstyrelsen för att snabbt kunna flytta kapacitet och tänk då på att vi har 21 stycken regioner och vi har säkert 200 sjukhus och massa annat och dem har olika system, olika sätt att definiera en och samma produkt så bara det kommer vara ett ganska hårt krav så småningom för att kunna ha en beredskap, att snabbt kunna få en blick över vart saker och ting finns precis som inom militären att man vill ha en snabb överblick och då måste det vara ett standardiserat sätt att hantera det.

Intervjuare (CD): **Du skulle alltså säga att det handlar mer om organisatoriska begränsningar?**

Respondent: Ja, det skulle jag säga, det är väl den största utmaningen och det säger jag också för att jag inte kan alla applikationer, det finns säkert massa

begränsningar. Det jag skulle vilja säga är att i denna förändrings-världen vi lever i så är den här omställningen hur man tänker och hur man gör, spelar ingen roll om det är utvecklare, revisorer, politiker eller företagsledare osv. Man säger att kulturen sitter i väggarna men det sitter i folks huvuden och vad man är van vid. Jag har jobbat länge inom IT-industrin och jag vet att jag själv har mina egna förutfattade meningar om saker och ting vilket kan hindra mig från att ta rätt beslut i vissa situationer, just detta tror jag är en väldigt stor sak.

[42:38] **Framtiden**

Intervjuare (RF): **Om vi kollar framåt igen, du har redan svarat lite på det skulle jag säga, men vilka är de kortsiktiga utmaningarna för er kring hanteringen av Big Data?**

Respondent: Jag tror lite grann att, ungefär som jag svarat tidigare, de tre kommande åren för oss alla företag, organisationer och verksamheter, är helt avgörande för vår framgång och då handlar det om just den här gemensamma synen på vad och hur vi ska transformera oss. Behovet av att all ska få en gemensam syn på vad Big Data innebär och inte bara som ett ord eller en möjlighet utan utifrån ett tekniskt perspektiv också, att alla måste bli lite Data Engineer. Man måste alltså vara på en liten nivå och börjar tänka utanför boxen. Som exempel, jag hade förmånen att få prata med en före detta generaldirektör för den danska digitaliseringsmyndigheten. Danskarna bestämde för nästan 15 år sedan att dem skulle digitalisera Danmark men att inte varje enskild kommun, landsting eller statlig myndighet själva skulle ha en digitaliseringsstrategi utan man skulle ha en för Danmark. Det första man gjorde var att man införde digitala brevlådor och all myndighetspost digitaliserades, allt från läkarbesök till deklarationer började gå digitalt vilket gjorde att men helt plötsligt inte behövde danska posten längre utan sålde den till svenska posten och bildade postnord och blev av med problemet med massa brevbärare och annat som kostar pengar. När de gjort det dock och kommit över tröskeln, att digitala brev är normalt, frågar man sig vad kan nästa steg vara och då tänker de så här, som exempel, behöver vi fysiska universitet? Lunds universitet skulle ha världens bästa föreläsare och kurser men att det skedde digitalt, man behöver inga fysiska byggnader eller fast anställd personal, samma sak med deras försvar, man kanske inte behöver ha massa marina fartyg eller stridsflyg osv. för att om någon vill invadera Danmark så behöver man bara slå ut elförsörjning, matförsörjning och banksystemet så skulle hela samhället ge sig. Det jag vill beskriva med detta är att om man väl klivit över tröskeln och tittar på vad man faktiskt kan göra, ungefär som Uber, behöver man taxibilar och taxiväxlar och chaufförer, nej det behöver vi inte för vi kan göra på ett annat sätt, man behöver komma över den tröskeln och förstå lite grann vad tekniken kan ge med ändrade affärsmodeller och sätt att bedriva verksamhet, då kommer det smälla till och de som kommer över tröskeln först oavsett vilken bransch det är kommer vinna. De som inte tycker det här är viktigt och bara tycker det är en fluga: “Vi kommer kunna ha kvar våra fysiska datacenter” osv. de kommer inte att vara vinnare. Kortsiktigt

så handlar det alltså om att förstå det här och jag brukar säga med tanke på min erfarenhet att till för tre år sedan så var det oftast att: "Det som du vill ha kära kund, det kommer i nästa version av den här applikationen" men nu är det tvärtom, det vi pratar om är digitala affärsmodeller. Tekniken har funnits i 4–5 år redan men det är verksamheterna som ligger efter nu och tekniken står inte stilla utan den utvecklas exponentiellt hela tiden så de som inte hänger på och kommer över den här tröskeln de har en gigantisk uppförsbacke att gå.

Intervjuare (CD): **Vi är ganska nöjda där från vår sida men är det något du tycker att vi missat att ta upp?**

Respondent: En sak som jag funderar på som kanske är lite utanför ämnet, det jag springer på även fast jag är till åren kommen bland folk som är lika gamla men även yngre är att man tror att den här digitala förändringen kommer av att unga människor som ni redan vet hur man ska förändra samhället och på vilket sätt och jag tror att det är ett missförstånd för att det ni pluggar är superviktigt för att ni själva ska få insikten av vad som finns bakom en applikation som ni använde dagligen eller ett spel osv. Det tror jag också är en viktig aspekt, att det inte den äldre generationen ska tro att den yngre generationen som vet allt och ska göra allt utan att det även är den äldre generation som jag som måste fatta det jag försökt att säga, att inte Big Data är något som de som är födda på 90-talet och framåt ska sköta, jag tror det är en missuppfattning och att alla måste tänka på att oavsett vilken ålder man är i så måste man förstå de här sakerna som jag försökt förklara för er.

[48:50] **Avslutning**

Intervjuare (CD): **Då vill jag bara signalera att transkriptionen nu avslutas.**

Appendix 4: Intervju 3

Intervjuare: Christian Dahlberg (CD), Rikard Funck (RF)
Respondent: Pia Carlsson .
Organisation: SEB
Datum, tid och plats: 21 april 2020, 14.00, Lund (Sverige)
Intervjutyp: Digital intervju över Microsoft Teams
Tid: 30:45

[00:00]	**Introduktion**

Intervjuare (CD): **Då sätter vi igång transkriberingen nu!**

Respondent: Ja, okej, kommer jag bevaras anonym eller hur tänker ni göra där?

Intervjuare (CD): **Det får du bestämma helt och hållet, även om din organisation vill behållas anonym eller inte också. Vi kan göra så att du kan få gå igenom intervjun och så får du i slutet berätta hur du vill göra?**

Respondent: Ja, okej, det låter jättebra!

Intervjuare (CD): **Då sätter vi igång!**

[00:34]	**Bakgrund och Big Data management**

Intervjuare (CD): **Vi kan helt enkelt börja med lite bakgrund kring dig, vem du är, vad din arbetsroll innebär, och hur länge du har jobbat på SEB?**

Respondent: Ja, jag började på SEB 1997, så det är ganska länge sen. Började jobba på tekniksidan som systemprogrammerare, om ni vet vad det är. Ja, ansvarade för transaktionssystem, IMS (*Information Management Systems*). Mest med stordatorsystem. Sen gick jag över och började jobba med de olika utvecklingsmiljöerna i och med att vi har ganska många utvecklare, så då har vi skapat olika miljöer som man sitter och jobbar inom med dokument som stödjer hur man jobbar på SEB. Sen har jag främst jobbat med modellering, och nu jobbar jag med Information Governance. Alltså Data Governance, och de bitarna kring det. Håller nu på med att implementera hur vi inom SEB ska jobba med Big Data, hur vi ska göra analyser och saker kring det.

Intervjuare (RF): **Intressant, intressant. Vi pratar ju mycket om Big Data i den här studien, så för oss hade det varit intressant att veta hur Du skulle definiera Big Data.**

Respondent: Ja, jag tittade ju på er definition då och jag håller väl med er. För mig är det att man skala upp och ner vid behov, men så är det ju inte riktigt här på SEB i och med att vi har en egen miljö. Vi kör ju inte i någon cloud-miljö eller så, utan vi kör ju här lokalt så det blir ju inte samma skalning

med upp och ner då hela tiden men, det är ju mycket data som ska bearbetas och behandlas, tittas på, sorteras osv.

Intervjuare (RF): **Okej, men du gick in lite där på er behandling av data och så, skulle du kunna förklara bara en översikt på er generella datahantering vad gäller Big Data? De olika stegen och hur det ser ut för er på SEB?**

Respondent: Ja, alltså vi använder ju den är miljön då för att… Vi har ju runt 1200 system på banken, så jag jobbar väldigt silo-baserat på retail, market och de olika områdena. Vilket innebär att vi tidigare haft, ja kanske 40 olika kundsystem, som vi nu försöker få ihop till enbart ett kundsystem. Men vi har ju som sagt 1200 system som innehåller data som är intressant för olika. Så vi lägger ner dessa i vår *Data Lake* som vi kallar det, alltså den här Hadoop-miljön då, så lägger vi ner system här så ska den konsumera all data och göra analyser, API:er, rapporter och så. Vi håller nu på att "lägga ner flera saker i sjön" som vi säger för att utvinna värde från datan.

Intervjuare (CD): **Okej, där ser man, tänkte bara flika in med en liten generell fråga. Är det mest data som ni själva genererar eller hämtar ni in från andra?**

Respondent: Alltså, vi som bank genererar ju data, vi hämtar inte in så mycket data. Utan det är mest referensdata som vi hämtar in, exempelvis vilka kurser aktier och fonder står i och liknande saker. Vi köper ju in mycket data om landskoder och annat. Men i annat fall är det ju vi som skapar datan själva här. När vi tar an en kund så skapar man ju data som behövs för den kunden, och vilka produkter den kunden köper och så. Så vi skapar ju mest datan själva. Vi har ju inte börjat med att ta in Facebook eller dylikt ännu. Vi har fullt sjå att ta hand om vår egna data först, hehe.

[05:01] **IT-verktyg**

Intervjuare (RF): **Okej, om vi då kunde djupdyka lite mer på din datahantering. Alltså, specifikt kring vad ni använder för olika IT-verktyg för att hantera datan i de olika stegen av processen?**

Respondent: Ja, vi har ju Informatica som verktyg, vilket vi har jobbat med ganska länge. Det är ju de vi använder, först och främst, jag jobbar inte riktigt inom den delen, utan mer hur man ska beskriva datan

Intervjuare (CD): **Okej, ehm, skulle du då säga att ni använder, om du kunde gissa eftersom du inte är så insatt, standardiserade system som Microsoft Azure, eller har ni mer skräddarsydda system eller verktyg som ni själva utvecklat in-house?**

Respondent: Nej, nuförtiden försöker vi använda system på marknaden. Vi använder framförallt Informatica, vilka är ganska stora och dyra verktyg för stora företag.

Intervjuare (CD): **Okej, finns det verktyg inom Informatica som ni använder i de olika stegen av er datahantering?**

Respondent: Alltså, ja, de (Informatica) påstår ju det men det är ju inte riktigt så tycker jag.

Intervjuare (CD): **Okej, vad menar du då, känner du att det är en brist inom Informatica? Kan du utveckla?**

Respondent: Ja, alltså de har inte de *capabilities* som vi behöver, alltså de förmågor som vi under resans gång har upplevt att vi behöver. Vi började 2016 med Informatica och har lärt oss efterhand, och kommit underfund med att vi saknar vissa *capabilities* som exempelvis visualisering av färdigprocessade data, och även kunna göra *knowledge graphs*. Ja, men visualisera data-landskapet om man säger så.

Intervjuare (CD): **Okej, hur kommer det sig att ni exempelvis inte använder andra verktyg för att visualisera er data?**

Respondent: Det gör vi, vi använder ju Tableau för att visualisera själva datat efter vi har gjort analyser och så.

Intervjuare (CD): **Har ni då verktyg som ni använder som täcker hela processen av datahantering?**

Respondent: Nja, eller det jag tänker på nu… Jag jobbar ju med att beskriva vår data, och då vill man ju förstå vad för data vi tittar på, eftersom vi har 1200 system där man läser ner stora mängder data så måste man ju förstå vad det är för data. Det kan ju stå "datum" i flera tusen fält liksom. Vad är det då för datum? Man måste förstå sammanhanget. Och då vill vi bygga modeller för det, datamodeller, och det är där det brister lite. Vi har inga verktyg för att bygga dessa datamodeller. Vi vill kunna se specifika data i dess sammanhang. Då behöver vi ha en modell, och det är väl där det brister. Sen att kunna titta på datan i sig, genom Tableau som jag nämnde, det fungerar ju bra, men det svåra för oss är ju att veta vilken data vi ska titta på.

Intervjuare (CD): **Yes, okej. Vi har en fråga här där vi egentligen vill intervjupersonen rangordna sina använda IT-verktyg från de som de upplever ger organisationen minst till mest stöd kring er datahantering. Så för er blir det mer applicerbart att kanske rangordna hur väl de olika delarna av processen för hur ni hanterar Big Data fungerar?**

Respondent: Ja, jag förstår. Alltså inhämtningen av vår data kan jag inte riktigt svara på. Men sen jobbar vi ju väldigt mycket mer datakvalitet och profilering av data, och där tycker jag de verktygen vi använder fungerar väldigt bra, skulle jag säga, framförallt för att se datakvalitet. Sen kan vi ju även se hur dataflödet går genom vår pipeline, även transformeringen av data och sådana saker vilket verktygen sköter bra också. Men sen är det ju det att

kategorisera och beskriva data. Vi har en metadatakatalog som i dagsläget beskriver varenda kolumn i alla våra tabeller med ganska grova beskrivningar om det är persondata eller inte osv. Det verktyget hanterar ju det som det ska, men är inte så användarvänlig eftersom det är själva affären som ska göra det och inga IT-tekniker. Då blir det genast mycket svårare för de att hantera det verktyget eftersom de naturligt sätt inte är lika vana att arbeta med de som våra tekniker. Alltså inte så användarvänligt. Sen är det även svårt att söka, svårt att veta vad det är man hittar, det är alltså svårt att veta kontexten helt enkelt. Sen har vi även ett sorts glossary där vi lägger in affärstermer som ska beskriva vår data i dessa kolumner som exempelvis kundnummer, kundnummer ska alltid ha en viss struktur och alltid vara unika som ska kunna gå att refereras till var respektive kundnummer finns i databasen. Det fungerar ju också. Men sen när man kommer på nästa nivå och vill göra dessa datamodeller som vi vill göra för att identifiera vilken data vi ska använda och kunna visualisera den, det är då det uppstår problematik inom våra verktyg. Generellt sett tycker jag det tekniska fungerar bra, det har vi ju jobbat ganska mycket med, men det är när vi kommer upp till hur affären ska beskriva sin data och lite mer organisatoriska aspekterna som är problematiska.

Intervjuare (CD): **Ja, okej, precis. Tror du det fungerar så bra med de tekniska aspekterna i och med att ni använder standardiserade lösningar? Alltså, tror du det hade varit svårare om era verktyg och lösningar hade varit utvecklade och skräddarsydda in-house?**

Respondent: Det gör vi ju inte, det har vi ju slutat med. Vi har en policy inom SEB som säger "*buy before build*", men vi måste ju ändå göra vissa hembyggen. I vissa fall använder vi exempelvis Atlas, som är en datakatalog också, som vi måste bygga egna lösningar som kan vara kompatibla med dessa för att kunna läsa in de. Dessa måste alltså kopplas ihop till våra befintliga system och verktyg. Det fungerar inte riktigt ibland och så, men det är ju fortfarande teknik, det svåra är ju organisatoriska aspekterna med kunskap och förändringsarbete. Nya tankesätt och så.

Intervjuare (CD): **Ja, okej. I eran sån här "Data Lake" som du nämnde, är det för all er data eller enbart för er strukturerade data som är redo för processande och analys? Finns fler "Data Lakes" för olika sorters data? Hur fungerar det?**

Respondent: Ja, eller alltså vi har ju en Data Lake kan man säga. Men egentligen kanske vi inte ska kalla det för "Data Lake", utan *information platform*. Vi har ju *information warehouses* också. Vi har ju några olika plattformar, men det som har fokuserat på sen 2016 är ju vår Data Lake, och där ska vi ju läsa in ostrukturerade data också.

Intervjuare (RF): **Okej, okej. Då ska vi se, vi kan nog lika gärna hoppa in på begränsningar på en gång nu då.**

[16:11] Begränsningar

Intervjuare (CD): **Ja, men absolut. Du har ju sagt lite nu angående begränsningar inom Informatica och de, vad du vill kunna göra och så för dina datamodeller. Vi kan börja kortfattat bara hur du skulle definiera en begränsning, i sin rena form, av en mjukvara?**

Respondent: Hehe, ja… Ehm… Att man inte kan göra det man förväntar sig. Om man då tittar på andra verktyg så klarar de av det man vill göra, men inte det vi använder. Då känns ju det som en begränsning för oss.

Intervjuare (CD): **Ja men precis, och nu när du nämnt Informatica som ert huvudsakliga IT-verktyg, vad skulle effekten av det här vara på er organisation med tanke på den begränsningen?**

Respondent: Ehm, ja det blir ju svårt att förstå att man tittar på rätt data, när man ska koppla upp sig mot Tableau och göra sina analyser. Då vill man ju veta att det är rätt data man kollar på. Om man exempelvis ska göra en rapport, en legal report till kanske finansinspektionen så måste man ju veta att det är exakt rätt data så man inte hämtar data från fel ställe, då kan det verkligen gå illa. Så det svåra är ju att beskriva data på ett sätt så man förstår det i sitt sammanhang, och visualisera det för användaren.

Intervjuare (RF): **Yes, okej, intressant. Vad hindrar er organisation just nu från att i så fall byta från Informatica till något annat? Givet den solklara begränsningen ni upplever.**

Respondent: Pengar, hehe, det är jättedyrt! Då måste man ju verkligen vara på fötterna och förstå problemet grundligt, och vi är nog inte riktigt där än eftersom det fortfarande är ganska nytt för oss. Vi ligger ju i framkant men sen kommer resten av bankerna som knappt förstått vad de ska använda verktygen till. På så sätt är det ju väldigt svårt att bara byta system eller verktyg. Det måste ju komma krav från affären, det är ju dom som måste ställa kraven, de måste förstå att de har det behovet, och det har de inte förstått än. Så vi får ju fortsätta jobba med det här tills de förstår vad de behöver. Då kommer det säkert kunna lösa sig. Då kommer de förstå att vi behöver nya verktyg för det här problemet. Men vi kan liksom inte sitta och säga att vi behöver nya verktyg, för då är det inga som tror på oss, för de förstår inte varför. Att vi alltså behöver nya verktyg.

Intervjuare (RF): **Ja, okej. Så du skulle alltså nämna att kunskap är en form av begränsning för er nu? Som du nämner, att ni/de inte riktigt vet.**

Respondent: Mm, mm absolut. Det är framförallt mognadsgraden så skulle jag säga, inte direkt enbart kunskapen, utan mognadsgraden mot affären som är ett problem.

Intervjuare (RF): **Hm okej, alltså, säg, rent teoretiskt - du har ju redan nämnt lite brister med Informatica - men om vi ska fokusera på ett**

framtidsscenario där ni får en högre mängd av data in i systemet, hur kommer era verktyg ställa sig mot det?

Respondent: Ehm, alltså, jag tror nog systemet skulle klara av det, mer än så kan jag nog inte svara på, men, ehm, plattformen är ju uppbyggd så, att det ska vara skalbart, så det ska ju fungera med stora datamängder. Mer än så törs jag inte svara på.

Intervjuare (RF): **Okej, inte heller om det skulle komma en förväntan på högre datakvalité från er sida? Hur skulle era system hantera det?**

Respondent: Ja, alltså, det jobbar vi ju också med. Det ingår ju i vår Data Management koncept, eller Data Governance koncept, att vi ska ha kvalité på vår data. Och det har vi ju beslutat inom SEB vilka kvalitétsdimensioner vi ska ha. Så att då får man ju titta på datan och ser man fel då så bör ju dessa fel ändras i källsystemet. Så att nästa gång den sortens data kommer in så kommer den vara rätt, vi vill inte sitta och behöva rätta datan som kommer in i själva sjön (Data Lake), eller på informationsplattformen, utan det ska rättas i källsystemen.

Intervjuare (CD): **Okej, okej. Skulle du uppskatta att det snarare är operativa, organisatoriska aspekter som en begränsning till att ni inte har optimala processer för er hantering av Big Data? Eller är det mer tekniska aspekter inom era IT-verktyg?**

Respondent: Alltså tekniken, det är ju aldrig något problem, det är bara att göra. Det svårare är organisation och kultur, framförallt förändringsarbete. Det är det svåra i vårt fall. Framförallt inte för ett stort företag som vi är. Tekniken har ju aldrig riktigt varit "en issue" utan det är mer processer och såsom är den svåra saken. Små företag kanske inte har råd att köpa stora etablerade verktyg, så då kanske det är lättare att ha mer tekniska problem inom extrahering, analysering, och visualisering och så. Vi är ju ett stort företag så där brukar det inte vara ett problem rent ekonomiskt heller. Beror ju på lite!

Intervjuare (RF): **Okej, intressant! Men då tycker jag vi blickar framåt lite grann och kollar på hur det kanske ser ut i framtiden.**

[23:14] **Framtiden**

Intervjuare (RF): **Vilka är de kortsiktiga utmaningarna för er organisation inom hantering av Big Data och era verktyg?**

Respondent: Alltså, för mig, eller för vår avdelning, så är det ju att få affären att förstå vad de kan använda våra verktyg till, alla olika nya affärsmöjligheter, och helt enkelt att få de att förstå syftet till användningen av dessa verktyg. Mognadsgraden är ju så pass låg. Inom Data Governance jobbar vi ju också med data-ägare och data-stewards. Data-ägare är de som är ansvariga för all data och data-stewards är ju de som jobbar med själva modellerna och vår data. Dessa roller har ju ännu inte landat än, man

förstår inte riktigt vad man är ansvarig över, och så, så vi är väldigt tidiga i detta kan man säga. Någonstans måste man ju börja.

Intervjuare (RF): **Jo, men precis, precis. Varför tror du att det är en utmaning att få folk att inse värdet med det här?**

Respondent: Alltså man har ju fullt upp idag med sin arbetsuppgift, och sen så kommer något nytt som läggs på, man förstår inte riktigt varför. Ehm, och då ska man lägga tid på det vilket man inte har, men det kommer ju liksom från ledningen eftersom vi ska vara den datadriven bank. Det börjar ju komma, men det tar tid helt enkelt. Vi är en gigantisk organisation. En väldigt stor organisation, och det tar tid innan alla förstår vad saker och ting handlar om. Det är bara att inse att det är ett väldigt stort fartyg som man måste svänga på, vilket kommer ta tid.

Intervjuare (RF): **På vilket sätt… Säg att det skulle ta väldigt lång tid för er ledning att inse värden av det här. Vad kan då effekten av det vara på er organisation i det långa loppet?**

Respondent: Alltså, ledningen har ju förstått att det behövs så där är det ju inga problem - men de förstår inte riktigt hur vi ska införa det kan man väl säga. De har ju bara sagt "Nu ska vi göra det här", sen är det ju liksom många steg som ska följas efter det. Och om de inte gör detta så riskerar vi ju vår tillvaro som bank. Då kommer ju alla andra att gå om oss. Sen har vi ju alla regelverk som vi måste förhålla oss till och alla rapporter som vi måste göra. Vi måste liksom ha koll på vår data. Vi måste helt enkelt bara göra det.

Intervjuare (CD): **Yes, okej. Då ska vi se… Ja, vi kan ju fråga dig; är det något du tycker vi har missat att ta upp som du vill tillägga?**

Respondent: Nej, det tycker jag inte. Jag vet ju inte riktigt vad ni är ute efter så, men hoppas ni fått ut något av värde, hehe. Vi kan ju ta upp det med ostrukturerade data, det är ju absolut något att tänka på, vi har ju fullt sjå med att ta hand om våran så kallade strukturerade data som redan finns i våra system. Men även ostrukturerade data som bilder, och såna saker, det jobbar vi ju också med parallellt. Vi använder taggar och sånt för att kategorisera data. Men mer utöver det kan jag nog inte ta upp på rak arm, var vi ligger där, vi försöker främst jobba med den datan som vi har som vi vet vad det är för data, alltså vår egna data. Sen finns även ljudfiler, är ju också något vi jobbar med men som vi ändå inte jobbat med så mycket.

Intervjuare (CD): **Vad är det som hindrar er från att använda den ostrukturerade datan på det sättet ni vill använda den på?**

Respondent: Det är nog för vi inte vet vad datan är för något. Och det är väl det som egentligen är problemet; det är det egentligen ingen som pratar om, alla pratar om tekniken. Det här med att man ska tagga textfiler, man ska ha bilder, man ska skanna dom osv, det är ju ändå bara teknik - vilket är ganska lätt att jobba med egentligen. Det svåra är ju att tala om, och

beskriva data, vad är det för data liksom, att kunna beskriva data så man vet att det är rätt data man använder - det är i alla fall utmaningen hos oss. Och begränsningar inom verktygen vi använder är ju då, som jag gick in på lite tidigare, att man inte får hjälp med de olika modellerna som vi behöver inom *knowledge graphs* exempelvis. Innan har det ju varit hemsidor som man refererat till genom länkar, men nu länkar man specifikt till data. Exempelvis på hyperlänkar på hemsidor så länkar man till data som man då sätter ihop. Vi har alltså gått ner ett steg från att länka till olika hemsidor till att länka till den specifika datan. Då måste vi veta vilken data vi måste sätta ihop. Och då måste vi bygga modeller för att veta vilken data som sitter ihop med vilken. Exempelvis har en kund ett kundnummer, och måste alltid ha ett kundnummer och namn, och då bygger vi sådana här grafer (*knowledge graphs*) för att få datan att hänga ihop - så då måste vi ha rätt data från rätt ställe, och veta vilket system av alla 1200 som innehåller rätt data, om jag nu ska försöka beskriva det så gott det går, hehe.

Intervjuare (CD): **Okej men super!**

[30:21] **Avslutning**

Intervjuare (RF): **Då meddelar vi nu att vi avslutar transkriberingen!**

Appendix 5: Intervju 4

Intervjuare: Christian Dahlberg (CD), Rikard Funck (RF)
Respondent: Respondent B (Anonymt)
Organisation: Företag B (Anonymt)
Datum, tid och plats: 22 april 2020, 09.00, Lund (Sverige)
Intervjutyp: Digital intervju över Microsoft Teams
Tid: 43:29

[00:00] Introduktion, Bakgrund och Big Data management

Intervjuare (CD): **Då kan vi börja med lite generell information om dig, ditt företag och din arbetsroll? Även möjligtvis inkludera din egen definiering av Big Data?**

Respondent: Jag är ett par år över 50 och tänkte dra lite historia för er här - och har jobbat med information i nästan 30 år, med olika former och mängder av data. Jag kan ha en liten annorlunda syn än vad många andra yngre har. Jag har jobbat inom de flesta olika branscherna; telekom-branschen, städbranschen. Ehm, jag har jobbat inom dagligvaruhandel, inom banker osv. Så jag har träffat på mycket data på många olika sätt, och min definition av Big Data är egentligen utifrån två perspektiv; antingen ur bredden av data, hur brett är datasetet? Det behöver inte vara många rader men det kan vara brett, eller så är det smalt, men med väldigt många rader. Det är två olika sätt att se datat ur den komplexiteten den har. Det som är Big Data för, vad ska vi säga, en stor organisation som Astra Zeneca eller liknande, eller om den är liten i deras värld så kan den vara väldigt stor i någon annans värd (det kallas för rörfirma). Så egentligen blir data "Big Data" först när man har ett verktyg som inte klarar av vanliga data. Då blir det man hanterar Big Data plötsligt. Så, när man inte har verktyg eller kompetens att behandla ens data så handlar det om Big Data, hehe, det är så krasst jag ser på det. Det är ju då min definition. Och jag menar, förr i världen när Excel kunde ta 32 000 rader, väldigt många år sedan men ändå, det satte ju en begränsning vilket gjorde en tvungen att kolla på andra lösningar som att lägga in det i Access eller någon form av SQL-databas och fortsätta jobba med det där, det är ju där rapportverktygen kom ifrån. Sen så växer det ju hela tiden. Det är min definition av Big Data; när man får för mycket data, när datasettet blir för stort.

Intervjuare (CD): **Om man kollar på er generella datahantering, om man kollar på hela processen från extrahering av data till visualisering av färdigprocessad/-analyserade data, skulle du kunna förklara den lite överskådligt?**

Respondent: Eftersom Företag B är ett konsultbolag jobbar vi ju inom många olika företag, så skiljer det sig ju för alla, vi har ju inte bara ett sätt utan vi är ju skomakarens barn, vi har ju ingen ordning på det själva internt. Och alla konsultbolagen jag har jobbat på så är det ingen av de som har haft

ordning på det internt, utan man bygger det mot kunder. Så är det. Man har inte tid att fixa till det på hemmaplan. Generellt, så har det tidigare varit så att man hämtat information från egentligen existerande databaser och existerande filer, sen har man satt samman det här till en mängd information so man har kunnat analyserat. Jag skulle vilja säga att det gradvis har gått över till att ha kunnat söka på webbtjänster och läsa av det genom utvecklade API:er. Det här håller ju på nu att ta ett gigantiskt skutt i och med IoT, där det regnar in data. Komplexiteten och svårigheten av IoT data är att veta vilken data som är intressant. Om jag har ett flöde på en maskin som har tio sensorer och man plockar händelser från var och en sensor så kanske det bara är två av de sensorerna som ger någon mening att följa upp, och då får man Big Data i mängder. Men man kanske inte har någon glädje av att det är väldigt begränsat. Så då är har vi en utmaning. Tidigare var det ju så att de flesta läste in data i en databas och sen bearbetade datan där, med de nya molntjänsterna som finns, exempelvis Google, Amazon och Azure så tror jag att vi kommer se en bild där man i mångt och mycket lägger all information i någon form av Data Lake som gör att det blir tillgängligt för många, däremot behöver man ju även sätta metadata och restriktioner som säger att det är okej att det följer ett visst mönster och så. Därefter så gör man någon form av Machine Learning för att plocka fram, vilket ger en möjligheten... eller om man redan vet vad man ska få fram så kommer man förmodligen aggregera upp informationen, stoppa det i en databas och sen rapportera ut det. Vilka verktyg som används, är ju ett gäng kända varumärken, men där är ju också en mänsklig del i det hela. Om man tittar på informationen så kan man ju inte plocka in för många dimensioner, eller parametrar, en människa klarar av 6 dimensioner att vända datan på, så det är en sorts begränsning på hur man kan se det utåt. Generellt hos de kunderna som vi jobbar med i södra Sverige och i Danmark för den delen, är det QlikView eller Power BI som vi använder. Och givetvis Excel. Så, datat utåt är Excel, men processen ser ut att man plockar in data till en area där man temporärt lagrar den, sen har man en area där man städar datat, och visa städar och gör affärsregler också, sen har du också modellering på det där man bygger modeller för Machine Learning, sen sparar man undan den mängden, sen kan flera stycken olika använda datan, där man också har en säkerhetsmodell inkluderat.

[08:05] **IT-verktyg**

Intervjuare (CD): **Ja, okej, intressant! Först och främst bara, hur pass standardiserade verktyg är det ni använder?**

Respondent: Jag skulle vilja säga att vi använder komponenter. Alltså färdiga komponenter och det gör alla, de som säger att de inte gör det ljuger. Både större och mindre företag. Jag kan dra ett case, ett städbolag som vill förändra sin, vad ska man säga, verksamhet eller hur man ser på de, de har beslutat att plocka in sensorer som läser partikeltäthet i luften ute hos sina kunder och sätter upp sensorer för att samla in informationen, och sen får de fram rapporter som de kan sälja tillbaka till kunden. De här använder alla komponenter, eller ja, sensorn är ju specialtillverkad

för där har vi haft ett företag i Österrike som löst det, men sen lyfter de upp data till *Azure* och sen visualiserar vi det i Power BI. Databasen i Azure är ju en standardprodukt, men man konfigurerar det internt med regler och så, det är det som är skräddarsytt för varje företag men själva verktyget är standardiserat hos oss. Det är ju inte så många som får för sig att bygga en dator nuförtiden, så om ni tänker det på det viset… Det kan ha varit en del modeller som använder Machine Learning som kan vara skräddarsydda i och med mängden data, och den typen av data, och man gör beräkningar, men själva samla in och lagra-stegen… det finns så många standardiserade verktyg så det är som att uppfinna hjulet igen, då är man väldigt kreativ, men delarna emellan, som att skapa ett API, det kan vara utvecklat på det viset, även med regelverk, de är ju specialgjorda för respektive verksamhet.

Intervjuare (RF): **De här verktygen som du nämnde tidigare, om du skulle kunna rangordna de utifrån de stöd de ger er datahantering, kanske på en skala mellan 1 och 5?**

Respondent: Ja, när det gäller data som ska lagras så, lite beroende på vilken plattform man använder, låt säga att du använder. Vi kan rabbla 5: Unix/Linux, Windows, On-Premises, Azure, Google och Amazon. Tittar man på de två förstnämnda och On-Prem, så kan man likställa filhanteraren där som en Data Lake i någon av de andra. När man får så mycket, mycket data så tror jag en Data Lake, eller "file storage" som koncept, det är steg 1 i datahanteringen eftersom det är där man lagrar sin data efter den är extraherad. Det är där man laddar in i sin Data Lake. Sen finns de dom som har mindre antal data, där man väljer att läsa in sin data rätt in i sin databas, typ SQL Server, Oracle, DB2 osv. Det är väl egentligen de två alternativen är hur man lagrar all form av information. Du har ju bilder och sånt också. Vi har gjort ett case uppe på ett stålverk i Värmland, Uddeholm, där vi läser bilder på järn som smälts ner. Så har man en kamera, vilket är det enda sättet att se om det blir splitter-bildning i järnet, och kan då adressera det direkt. Det läser man ju inte in i en databas, men det läser man in i någon sorts *file storage* där man matchar bilderna mot ett... Ja, du har en Machine Learning algoritm där man läser av bilderna och säger "ja, okej här är en splitter-bildning, gör om och iterera". Där är ju bilder också, men det är ju egentligen bara ettor och nollor i ett flöde... Det är väl svaret på var man lagrar, det är ju som jag ser det bara två ställen, sen finns det ju olika sorters databaser som ni kanske vet; objekt-databaser och vanliga transaktionella databaser och relationsdatabaser. Där är väl det ju den transaktionella databasen som är störst. Sen har man ju ett antal olika integrationsverktyg, och vissa skriver sina egna applikationer i C# eller Java som hämtar data, andra låter några andra pusha data till dom, det är lite strukturen på organisationen däremellan som bestämmer. Sen rent generellt med Big Data så är det ju Machine Learning som är huvudfokus, vilket absolut inte är min stora styrka vill jag poängtera, men där finns ju en del olika verktyg som ML studio (Microsoft Azure Machine Learning Studio), som byter namn lite då och då, hehe, sen finns det även någonting hos Google, ett annat koncept som

heter BigQuery, som är lite annorlunda uppbyggt än ML studio men jag är inte så bra inom de delarna…

Intervjuare (CD): **Okej. Men jag tänkte mer… De här verktygen som du nämnt, tycker du, genom alla de olika stegen inom datahantering, de hjälper er som ni vill eller är det några fall där det inte går så bra som du hade önskat?**

Respondent: Nej, de som finns på on-prem (On-Premises) är välutvecklade. De som finns i molnet håller på att bli bättre, men de är inte lika komponent-starka som resten. Därav att det behövs många som kan programmera nu i det här teknikskiftet.

Intervjuare (CD): **Haha, ja okej, det är ju kul att veta för oss intresserade av programmering.**

[17:05] **Begränsningar**

Intervjuare (RF): **Ehm, vi tänkte gå in på lite specifika begränsningar med de verktygen som du har nämnt, och då ville vi först fråga, för det är kul att veta, hur du skulle definiera en "begränsning" inom en mjukvara?**

Respondent: Haha, ja, en mjukvara jag bygger eller en mjukvara jag skall använda?

Intervjuare (CD): **Både och nu när du lägger upp det så!**

Respondent: Ja, alltså, om jag bygger en mjukvara, vilket jag har gjort några gånger, så finns det en begränsning i saker som finns att använda. Vi byggde nämligen en inventeringsapplikation väldigt tidigt där man skulle inventera ett lager och där hade de skrivit direkt på skyltar vad det för någonting, där var den tekniska aspekten att man inte lyckades tolka riktigt vad som var skrivet, så man fick helt enkelt sätta upp en QR-kod. Så, det finns en teknisk begränsning när man gör en applikation; att förstå människans beteende, typ hur personen skriver och så. Det är människan som är stora boven, att vi inte tänker likadant allihop, och har olika värderingar, åsikter och vad vi tycker är viktigt och så. Det är ju givetvis charmen men det är ju det som sätter en form av begränsning. I de verktygen som vi har nu och använder så ser jag inte att vi har något direkt som saknas. Och om det är något som saknas så får man helt enkelt bygga det själv. Jag vet inte riktigt om det är svaret på er fråga, hehe.

Intervjuare (CD): **Jo då, du kom in lite på vår nästa fråga som handlar om vilka begränsningar ni upplever inom era IT-verktyg som ni använder inom er Big Data-hantering?**

Respondent: Ja, alltså det finns faktiskt två grejer som är viktiga, dels ur ett regulatoriskt perspektiv, det är att kunna följa data. I bankväsen så måste man ju veta var datat kommer ifrån, och måste ha spårbarhet på både informationen och på vilka kalkyleringar som har gjorts och så. Ni tänker

er en bank, så här fungerar det på banker nämligen; var och en bank har uppgifter att en gång i månaden lämna upp en satans massa information till EU för att bevisa hur de driver bank. Och då finns det ett antal data poäng där som de plockar ut, och utifrån dessa gör de då vissa beräkningar, så kan det gå fem månader, så kan EU-kommissionen komma tillbaka till banken och fråga "hur räknar ni ut detta månad 5 förra året?" då måste man ha lagrat allting, alla dataset, eller kunna spåra ur man gjort, för om man gör en ändring inom programmet så måste man kunna spinna tillbaka och göra om det programmet för att kunna hämta ut informationen. Det är ett sätt det blir Big Data, det är att myndigheterna kräver det. Den ska du spara i 10 år, eller något sådant.

Intervjuare (CD): **Du nämnde två begränsningar tidigt i det du sa, eller två saker, om du har några begränsningar inom era IT-verktyg, vad är då effekten av dessa på er organisation som helhet?**

Respondent: Alltså, på vår organisation så blir ju effekten där att kunderna upplever att det blir för dyrt. Att de tycker det ska fungera ändå. "Det är väl inte så svårt" tänker dom, men hela det här om att spara undan och kunna följa upp kräver ett helt annat tankesätt. Jo, de två grejerna jag tänkte på var *traceability* och *lineage,* lineage är att du ska kunna följa var du har varit någonstans, veta alla stegen man har gjort. Det är två begränsningar som jag upplever finns inom både molntjänster och on-premises. Vi använder hela Microsofts flotta av grejer, då är det lättare att följa de, men vissa företag har lite Oracle databaser och DB2, så kör de Informatica som verktyg, de gör att det blir svårt att följa sin data då. Så här tror jag det kommer bli en stor marknad eftersom man kommer vilja kunna följa sin data genom hela processen.

Intervjuare (CD): **Yes, okej. Gällande de här begränsningarna du har nämnt nu, hur appliceras de på standardiserade verktyg gentemot internt utvecklade verktyg? Finns det alltså anledning till att ni inte byter verktyg med dessa begränsningar i åtanke?**

Respondent: Ehm, det kommer nog att bli ett krav på de verktyg som plockas fram. De som finns nu, de verktyg som stödjer det, de är så fruktansvärt dyra. De går på flera miljoner att köpa in som kan stödja spårbarhet och så, och de är lite för mycket för många.

Intervjuare (RF): **Om vi då fortsätter med de IT-verktyg som ni använder, om det i framtiden skulle komma ett ökat behov av att hantera större mängd data, hur skulle era verktyg hantera det?**

Respondent: Ja, alltså de verktygen gör det. Det som är begränsningar är ju, när det gäller att skyffla data, är ju både fysiskt minne och processminne. Lineage-verktygen är ganska basala egentligen, för de pekar ut ett dataset som kommit från a och vandrat till b. Därför tror jag att lineage och traceability kommer bli ett krav. Man måste, när du sitter och tittar på Big Data, så måste man veta var datan kommer ifrån.

Intervjuare (CD): **Yes, okej. Du nämnde processkraft där, är det något du upplever nu? Att era verktyg har mindre av det, eller är det enbart ifall det skulle uppstå ett behov av att hantera större mängd data?**

Respondent: Nej, det som är sköna med molnet, vilket är nästan bara där jag jobbar nu och har gjort under det senaste året, är att det är skalbart. Det är som att vrida på element-knappen, det kostar mer men man får mer utrymme exempelvis. Och då är det ju Azure vi använder. AWS är väldigt få här nere i södra Sverige som använder. Tittar du på företagen generellt i det här området så har de en viss storlek, alltså en storlek på sina databaser och liknande, som Microsoft erbjuder, är tilltalande för dom. Så då skriver de ett avtal så får de subscriptions ganska billigt så är de igång. Det som är fördelen med Microsoft är att de har allt genom hela processen, och framförallt att de har Power BI som slutverktyg för visualisering osv. De andra har inga sådana. I alla fall inte så uttalade.

Intervjuare (CD): **Nej, okej. Vi var inne där lite på hur era verktyg skulle hantera en ökad mängd data, och så, hur skulle det se ut ifall det istället skulle handla om ökad förväntning på datakvalité?**

Respondent: Ja, det är ju lite så, det man stoppar in i verktyget kommer ju också ut ur andra änden. Många gånger är det så att källsystemen som levererar innehåller ganska så mycket brister, även om man inte tror det. Det är väldigt sällan det hänger ihop. Detta gör att resultatet utåt tror man ska se ut på ett sätt, men visar sig vara något annat. Har du till exempel ett pris på en produkt som är satt att det gäller från ett datum och framåt, sen så råkar det finnas ett annat prisintervall som det överlappar, så går man då på det datum där de överlappar så kommer det bli dubbla rader. Den kontrollen i många interna system finns inte. Och dessutom så läser man in det här, i mångt o mycket, i Excel vilket gör att det är många som fuskat och sätter samman och använder former där. Datakvalité är ett stort bekymmer efter systemen inte fixar det. Det måste man alltid rätta upp om man vill ha ut något schysst där man samlar information.

Intervjuare (CD): **Ja, okej, det blir lite dubbeljobb alltså?**

Respondent: Ja, precis. Du måste ju liksom gå tillbaka och se om du ska rätta ett fel eller bygga någon "regel" vid sånt fall, fram och tillbaka med flera upprepningar. Försöka täcka alla olika specialfall. Det är ju också en anledning, som jag tror, när man får in så mycket data, innan har alla dessa uppföljnings-systemen varit hundraprocentiga; det skulle vara exakt rätt. Jag tror att man kommer nöja sig med en, i många fall, 95% rätt, utan man har mer *sanity check* där man säger till sig själv "ja, det bör ligga här någonstans", det kommer kosta för mycket att rätta allt varje gång. Man kanske får spendera några miljoner hit och dit, det är i vissa fall skitsamma om man pratar om miljarder, men det är en begränsning i källsystemen i alla fall, att man måste rätta fel i efterhand, som bidrar till dubbelarbete och en ökad kostnad.

[32:22] Framtiden

Intervjuare (RF): **Utifrån det du har sagt med olika begränsningar på olika verktyg, vad är de kortsiktiga utmaningarna för er och er Big Data-hantering?**

Respondent: Jag skulle vilja säga, eller tro, att organisationer inte riktigt vet om att de kan få en Big Data lösning, och vad de ska göra med den. De har liksom ingen plan för att, alltså, de kan ha en kortsiktig plan men ingen långsiktig plan. Det är det jag känner, man är väldigt omogen i det här, vilket gäller inom alla branscher, sen är det väldigt många som har löst tre år, jobbat ett år som vill bli Data Scientists. Det är för mig väldigt tufft, det krävs rätt mycket, det är någon form av glamour-värld där man tror att jobba som Data Scientist. De som läser statistik ihop med er, de är dom som kan hitta sambanden i information. Där tror jag att vi saknar en del förmåga inom företagen.

Intervjuare (RF): **Du pratade om den här omognaden, vad tror du den beror på?**

Respondent: Ekonomer. Omogenhet beror på att det är ekonomer som styr och ställer. En ekonom tittar alltid bakåt, och säger "det här måste vi strunta i för det gick inte bra", istället för att säga "det här gick inte så bra, vi kanske borde satsa på det ännu mer så att det kommer gå bra". Ekonomer är destruktiva människor i det fallet. Eftersom ekonomer styr många företag, och styrs mycket genom att tjäna pengar, istället för att styras för att göra saker bättre och effektivisera världen. Vi slipper jobba, vi slipper göra det, så kommer det ju kosta en del, men jag menar, hela den här Corona-situationen måste ju vara en mardröm för alla ekonomer. Vi inom tech sitter hemma, det här funkar ju jättebra (remote-intervjun), det hade inte funkat för tre månader sen. Hade jag sagt att vi skulle göra det här så hade ni sagt "han är ju inte klok". Så, begränsning av att vi inte kommer längre är, dels ekonomernas fel, men sen tror jag att det finns ett ointresse för detaljer. Jag kan berätta om ett case, vet ni vad Grundfors är? De gör pumpar för oljeplattformar eller liknande. Detta är ett sådant Big Data och Machine Learning case där de upplevde att det va för dyrt att plocka upp pumpar när de gick sönder, och oftast var det samma sak som gick sönder, men att köra ett antal Machine Learning-algoritmer så märkte man att 83% av alla pumpar under produktion var defekta som man kunde ta bort. Det är först när man berättar det för en ekonom, "vet du vad, det här projektet kommer kosta en miljon", då svarar de "nej, men då kör vi inte det, alldeles för stor kostnad", sen fortsätter man "men du kommer spara 100 miljoner i längden", då ändrar de sig snabbt till att vilja köra på det. Inte att "vi kör projektet för att minska utsläppet, minska transporten, minska lidande", det är ju inte det de handlar om, utan pengar ut pengar in.

Intervjuare (CD): **Det du har nämnt nu om kunskap och sådant, skulle du uppskatta att det är mer operativa, organisatoriska aspekter som är problemen bakom begränsningar och bristningarna eller är det mer tekniska?**

Respondent: Nej, jag upplever att det är organisatoriska aspekter; man är inte där ännu. Dels är det ju, jag är ju gammal person (50+) egentligen, men de som sitter och styr som är ännu äldre, och inte ser det här... Dels har man svårt att se komplexiteten och viljan för detalj, det är för jobbigt att ta till sig tror jag. Det är självklart personlighet och sådant, man har en verksamhet som går bra och så, skruva inte i något som fungerar liksom. Det brukar vara bra, funkar det så låter man det vara, men de tänker inte på att det kan funka bättre… Då vill man ju inte röra det heller om det väl funkar. Så jag tror det mesta är organisatoriskt.

Intervjuare (CD): **Ja. Okej. Om man då kollar i det långa loppet, hur tror du att de begränsningarna du nämnt kan påverka er på lång sikt?**

Respondent: Jag tror att det är många som måste bli bevisade först. Det caset vi gör för städfirman jag nämnde tidigare, det kan ju vara en ögonöppnare för många i den branschen, att man kan samla information och sen basera på partikeltätheten i luften, att det inte har städats så bra, som visar att man måste städa ofta och kanske annorlunda. Eller att kanske luftflödet i huset är fel, att man måste byta fläktar eller så. Vi kanske inte ska ha ingång till butiken på den sidan för där dras det in fler skit och så… Det finns massvis att göra med information, men du måste vara intresserad för att hitta de grejerna. Det går ju inte att tro att allt kommer till dig, du måste ha ett intresse för detalj. När personer blir intresserade för detaljer, sen vill man ju inte snöa in sig, men då kan man hitta samband där man kan effektivisera saker. Jag vet inte, man gör ju själv hemma att man ställer tvättkorgen nära tvättmaskinen, av anledning till att det är lättare att hiva in det lättare, det har ju vi effektiviserat som människa. Det handlar om att hitta dessa sätt, och då finns det data till hjälp för att se. Finns ju Space Management inom konsumenthandeln och så för att veta var man ska ställa sina varor för att höja försäljningen, som att mjölken står längst in för då måste man passera allt innan man når dit och kanske plockar på sig lite på vägen. Men det finns ju också undersökningar att man sätter upp öl och blöjor på fredagar, eftersom det har visat sig att det är mest män som går och handlar då när frugan är hemma och haft jobbigt på dagen. Det är en gammal, gammal undersökning.

Intervjuare (CD): **Yes, okej… Ehm, ja men det känns som vi faktiskt fått svar på mer än vi hade frågat om. Finns det något du vill tillägga som du anser att vi missat gällande hela det här ämnet?**

Respondent: Nej, jag tycker ni har fångat det, och det är ni som ska visa upp någonting, hehe, så jag tycker det har varit intressant och prata med er. Jag skulle gärna vilja ha ett uppföljningsmöte sen när er studie är klar så kanske jag kan ge er feedback på den om ni vill.

Intervjuare (CD): **Ja men absolut, du kan gärna få ta del av hela studien om du vill!**

Intervjuare (RF): **Ja men precis, det är alltid välkommet med lite reviews.**

[37:05]	**Avslutning**

Intervjuare (CD): **Vill du bevaras anonym? Det missade vi i början.**

Respondent: Jo, men gärna det. Anonym från ett hyggligt stort bolag. Det underlättar.

Intervjuare (RF): **Inga problem! Det löser vi.**

Intervjuare (CD): **Då vill jag bara signalera att transkriptionen nu avslutas.**

Appendix 6: Intervju 5

Intervjuare: Christian Dahlberg (CD), Rikard Funck (RF)
Respondent: Respondent A (Anonym)
Organisation: Företag A
Datum, tid och plats: 24 april, 09.00, Lund (Sverige)
Intervjutyp: Digital intervju över Discord
Tid: 44.46

[00:00]	**Introduktion**

Intervjuare (CD): **Alright då kör vi igång! Vill du vara anonym eller ej?**

Respondent: Jag väljer att vara anonym.

[00:41]	**Bakgrund och Big Data management**

Intervjuare (CD): **Okej perfekt! Vi kan börja med att du får introducera dig själv, vem är du? Var har du för roll på företaget och hur länge har du jobbat där?**

Respondent: Jag har jobbat med lite olika roller men primärt som Solution Architect eller Application Architect beroende på vad det är för typ av projekt eller lösning man tar fram. Jag har jobbat både i linje samt i rena projekt rörande Data Warehousing. Jag vill börja med att poängtera att Big Data är inget nytt för mig, Big Data har alltid funnits, bara det att man myntade begreppet 2005, dem problemen som man definierar inom Big Data har alltid funnits. Jag har jobbat inom Big Data i många år, mer än 15 år, dvs att man kan lita på datan, sourca den och göra den tillgänglig osv.

Intervjuare (RF): **Skulle du kunna förklara lite översiktligt runt eran allmänna hantering av Big Data**

Respondent: Yes! Det beror först och främst på vad det är för typ av data, är det strukturerad data, semi-strukturerad eller ostrukturerad osv. Man har inte en "*one size fits all*" lösning, det närmsta man kan komma, och nu pratar jag från personlig erfarenhet, jag kommer inte generalisera allt för mycket, vi har ju gått data lake approachen där vi helt enkelt ser till att vi sourcar allt till ett och samma ställe och bygger då kontroller när vi injectar datan till data laken som ett första steg så sätter vi ett antal kontroller eller capabilities på den där för att uppnå de här olika sakerna vi behöver uppnå, att vi ska kunna lite på datan. Vi ska kunna tagga datan för att man senare ska kunna få ett värde av den för slänger man bara all data i en hög kommer ingen kunna använda den och ingen kommer kunna hitta någonting i det. Det är den viktigaste grejen, det måste medföra värde, annars är det bara nonsens för det är så många Big Data-projekt som bara går ut på att samla stora mängder datavolymer och sedan kan de inte göra något med det. Där är det också viktigt att det är ju inte alltid

man vet att man måste titta på fullständighet av data också. Ibland har du ett väldigt brett dataset t.ex. eller väldigt mycket attribut på ett dataset, det är möjligt att bara fem av fem hundra är relevanta i nuläget men det betyder inte att man ska skala bort 495 utan man tar in det och sen ser man till att man kontinuerligt kan uppdatera den här informationen och då använder man sig av förhållandevis lätta grejer som crawlers eller data officers som ser till att man uppdaterar och taggar informationen så gott det går, även ML-modeller kan man hjälpligt använda för att tagga information. Alla de här grejerna jag nämner nu är sådana saker som jag har erfarenheter av att använda, sen finns det ju många fler saker än så, det finns ju ”*off the shelf*”-produkter, jag är inte jätteglad för dem, dem är oftast väldigt specifika och klara en grej eller så är dem hiskeligt dyra och klarar lite mer men ändå inte allt så man måste ändå bygga på något själv eller ha multipla lösningar. Data lake konceptet är ju inte heller egentligen nytt, det finns i princip sedan tidigare i form av Data Vault, man lägger datan i en temp-area så rått som möjligt innan man skjuter vidare den och applicerar business rules.

Intervjuare (CD): **Jag tänkte på det du nämnde kring att ni har Data Officers som kontrollerar datan, är det större mängd manuellt arbete eller är det även ML?**

Respondent: ML! Det är output från en ML-modell. Primärt är de det. Vi har ju vissa flöden direkt från ERP-system dvs. system där datakvalitén är väldigt hög, man vet vad det är för typ av information och då kommer de här flödena vara IT-governered så att där är det inte så mycket att göra och där kan vi, som en del av extraktet från ERP och injektion eller import i data laken som en del av den integrationen, tagga de här dataseten med relevant metadata. Denna information lagras i ett metastore som sedan blir tillgängligt och sökbart för andra system dvs. System-to-System men även för Data Scientist, Business Analyst som då kanske behöver hitta ett specifikt dataset eller en specifik typ av data, då kan dem söka på kategorier eller nyckelord och få upp relevant metadata kring de här dataseten. Detta funkar ju primärt för tabulära dataset men det fungerar även för, till viss del, ostrukturerade data så länge man taggar informationen men då kanske man inte får upp så mycket relevant strukturell information om det men man får åtminstone en övergripande bild kring vilken information det är samt vilken avsedd målgrupp det är för den här typen av information. Sen har vi den andra grejen vilket är när vi har information som kanske kommer strömmande, det kan vara chatinformation, positioneringsinformation, lite vad som helst, den är ju lite mer opålitlig eftersom den är beroende av att en maskin i vissa lägen ska rapportera sin t.ex. position och den är ju mer eller mindre exakt så att där är det lite lurigt, då måste man på något sätt tagga att informationen är opålitlig eller ofullständig i vissa lägen, det är också attribut man måste hantera på datan. Ett annat exempel är om du är beroende av att en enhet ska skicka uppdateringar vid visst intervall så kan den enheten missa på grund av att den inte har täckning, som exempel, då blir det ett glapp i informationen och då har man ML-modellerna som kommer in och tittar på discrepancy i flödet och taggar

upp det och sen får någon då godkänna beroende på hur viktigt det här datasetet är och även titta så att det inte kommer med, speciellt när det gäller, någon känslig information, då måste man skrubba det, som även är viktigt när det gäller positionering på enheter, man får inte lov att skicka med något som identifierar enheten, det är många kontroller och det finns inget verktyg på marknaden som gör detta utan man måste bygga in de här grejerna. Lineage är också viktigt, man måste veta var den kommer ifrån, vad som händer med den på vägen osv.

[11:15] IT-verktyg

Intervjuare (RF): **Kan du, utifrån dina erfarenheter, berätta liter mer om de IT-verktyg ni använder för att hantera stora mängder data (Big Data)?**

Respondent: Absolut! När det gäller integration dvs. Hantering av de här dataflödena då jobbar jag primärt med saker inom AWS och där finns det ett antal verktyg, du har t. ex Glue för viss integration av data även då som crawlar för att vi ska skanna dataseten och hitta anomalier och identifiera känslig information osv. Vi använder oss av DynamoDB bland annat som metastore för att hantera metadata kring datan, göra det sökbart osv. För att vi ska få någon form av Governance på plats och spårbarhet av datan. S3 för att rå-lagra själva informationen. Vi använder oss av parquet, filformatet, för att faktiskt lagra och göra de här dataseten relativt snabba, när det gäller att ställa frågor och liknande. Vi använder Amazon Redshift som både klassiskt Data Warehouse men även som en ingångspunkt ner tillbaka till Data Laken via Redshift-spektrum, vilket är en extension för att man ska kunna köra så kallad schema on read mot data laken. Vi har även delvis använt Collibra. Collibra är ett data lineage verktyg. Data lineage är en väldigt viktig del av overall Governance men är också extremt svårt att få till för att man vill ha det så granulärt som möjligt men samtidigt inte för att det kan hända väldigt mycket med data på vägen från källa till alla möjliga olika rapporter och analyser och ML-modeller osv. Samtidigt har du även att en ML-modell kan spotta tillbaka ett nytt dataset baserat på existerande dataset och sedan kan den iterera fram och tillbaka, det kan hända väldigt mycket på vägen. Om jag ska säga någonting som jag saknar generellt när det gäller Big Data i allmänhet så är det data lineage, dvs. Spårbarhet. Man måste kunna spåra informationen man måste kunna förstå vad som har hänt med den för att kunna lite på den för att har man inte det så finns alltid risken att du har fem olika dataset som egentligen är samma sak då blir det Data Sprawl och det är det som är ett av de största problemen, inte bara att du får en väldig massa data utan att du har samma data överallt.

Intervjuare (CD): **Du menar alltså att det blir redundans?**

Respondent: Redundans är inte dåligt utan det beror på, låt oss säga, om du har ett dataset och gör ett aggregat av den, en gruppering baserat på vissa saker och sedan trycker tillbaka detta då är det inte redundans i min värld för det kan vara att båda dem två tillsammans kan vara relevanta speciellt ur

ett ML-perspektiv där du kanske vill verifiera någonting, då är det bra att ha de här, för då har du en point in time bland annat.

Intervjuare (CD): **Det kan också fungera som en säkerhet ifall data går förlorad. Jag vet att AWS använder sig mycket av detta i sina data warehouses.**

Respondent: AWS har ju väldigt mycket redundans. De har du i alla steg, men jag tänker mer om man tänker sig redundans ur ett business-perspektiv. Redundansen kommer du ha, kör du cloud har du redundans, du kan även ha geografisk redundans vilket man i vissa fall måste ha på grund av diverse lagar om integritet.

Intervjuare (CD): **Det känns som ni mest använder standardiserade lösningar men du nämnde även att ni måste ha en del in-house byggda lösningar för mindre processer.**

Respondent: Absolut! Då använder vi så kallade lambda-funktioner primärt för att det ju de olika checkpoint-valideringarna. När data kommer in måste man först och främst kolla på, är detta här ett nytt dataset? Dvs. skapar man ett nytt dataset måste man bygga upp ett schema kring datasetet, en struktur och schema med capabilities och det ju Lambda som gör detta, den parsear informationen som kommer in, är det en JSON-fil, är det en flatfile, att tillägga är att jag syftar på AWS-lambda, jag håller mig inom AWS-sfären för det är den största lösningen och den är den mest kompletta lösningen. Lambda är *shortlived* vilket gör dem perfekta för sådana här saker som sen gör de här checkpoint valideringarna, först och främst tittar de på strukturen på informationen gör ett förslag och sedan valideras själva datan mot detta. Vi måste alltså ha capabilities på informationen, däribland datatyper osv. Det är primärt den typen av grejer vi använder, vi använder ju standardiserad funktionalitet men gör det till vår egen, det måste man göra för det finns inget som löser alla problem, i alla fall inte i vårt fall.

Intervjuare (CD): **Hur väl anser du att de verktyg ni använder för hantering av Big Data ger stöd för det som ni behöver dessa för? Hade du kunnat rangordna hur bra stöd verktygen ger i varje steg i er Big Data-hantering?**

Respondent: Överlag har det fungerat bra i samtliga steg. Jag kan säga som så att när det gäller IT-govern dvs. när vi behöver flytta information från on-prem eller annat cloud, beroende på vad man har för hosting på våra stora källsystem dvs ERP, CRM, den typen, så har vi standardiserade integrationsplattformar sedan tidigare, där använder vi oss av Informatica cloud och Informatica on-prem men cloud primärt. Den integrationen där emellan fungerar bra men det har också att göra med att vi har en dedikerad lina från våra datacenter upp till AWS datacenter, vi har helt enkelt dedikerad kapacitet. Företaget jag suttit på har bra ekonomiska muskler så att det är inga problem, det är inget som de flesta företag har råd med så att den biten har fungerat smärtfritt gällande de datavolymerna som vi hämtar från on-prem eller cloud till cloud, det är

inga jättevolymer utan våra volymer kommer från redan existerande cloudtjänster och de ligger då oftast i Amazon också så det är egentligen bara transfer data mellan olika tenants i Amazon. Det har inte varit något problem men det kan vara ett problem, när vi gjorde våra test-cases i början innan vi hade allt på plats så var det ett problem så jag kan tänka mig att det skulle kunna vara ett problem i de flesta fall.

Intervjuare (RF): **Vad är det för problem mer specifikt?**

Respondent: Exempelvis latency, att man chokar linan, om du har många integrationer som går samtidigt, du måste pussla väldigt mycket för att få till det för det kommer gå genom den publika linan ut antingen kapslas via VPN eller via IP-whitelisting men det kommer fortfarande gå genom den publika linan så det påverkar, t.ex. om du har en proxy eller liknande mellan så påverkar det alla. Så det är där skon klämmer medan vi isolerar det till en egen lina och då blir det inte ett problem.

[23:01] **Begränsningar**

Intervjuare (RF): **Hur skulle du definiera en begränsning inom en mjukvara?**

Respondent: Den löser inte problemet jag vill ha löst. När man har ett Big Data-problem, som jag ser det, så är det för att nuvarande plattform inte klarar av att leverera information i tid. Jag såg ni använde den här 5 v definitionen, det finns ju allt från 3-7 v, jag gillar dem mer eller mindre men en grej jag reagerade på är att i er definition, velocity, ni pratar om frekvensen av att data ska läsas in men en viktig faktor här är också hur snabbt måste man hinna parsea och processa informationen för att göra den tillgänglig till en process eller funktion så det är två delar på den (velocity) som man måste ta hänsyn till, många gånger är det inga problem att antingen parsea datan och trycka in den och lagra den någonstans, alltså bara trycka ner den i en data lake, det är kanske desto svårare att få den tillgänglig i en slutrapport eller vad det kan vara så det är den som brukar vara det största problemet (velocity). Det är inte problem att läsa in det och ta hand om stora volymer ytan du ska hinna processa den och applicera businessregler och göra den tillgänglig. Jag tror också att det var veracity, som jag ser det så är det väl mer hur 1: Hur oföränderlig informationen är, den ska vara pålitlig den ska ha samma struktur, den ska inte avvika men även hur komplett den är, det går tillbaka lite det jag sa innan: Har du ett dataset med 500 datapunkter/attribut men du väljer att bara använda 5 av dem och bara ta med 5 av dem, värdet på den datan kan vara jättehögt men den är inte komplett för du vet ju inte om någon av de andra datapunkterna är relevanta för de 5 första som ni har identifierat, detta är också en jätteviktig grej, hur komplett informationen är som man tar in och lagrar. || Det är många gånger som folk tar den snabba vägen [relaterat till det han säger kring 500 datapunkter och bara ta med 5] och säger “nej men vi förstår inte vad de här andra grejerna är och vi tar bort dem för det tar bara plats”, ja, men det är omöjligt att lägga tillbaka dem senare.

Intervjuare (CD): **Skulle du kunna beskriva lite kring begränsningarna du upplever inom de verktyg du nämnt tidigare, om där finns några?**

Respondent: Lambda är ju som jag sa, där har vi de största problemen för att lambda är shortlived dvs. De lever bara under en viss tidsrymd och kan bara exekvera under denna. Vi har fått göra piggie-back på en massa lambdas för att lösa vissa problem. I början kunde vi lösa allting under en viss tidsrymd med vi har ju sprungit på problemen att den inte räcker till och då får vi timeout. Det är den största grejen men det var ju också ett medvetet val att använda Lambda, det finns ju annat man kan använda men vi valde att använda det för att vi vill köra det så serverless som möjligt och verkligen köra en *pay-as-you* go arkitektur, vi var med medvetna om att begränsningen kunde slå till vilket den också har gjort vilket absolut är en begränsning, att man är styrd av det. DynamoDB och Glue har också sina drawbacks, framförallt när det gäller accesskontroller och accessisolering är Glue inte jättebra, men har blivit bättre. Det är något som vi har haft problem med, eller snarare problem med att få godkänt av diverse security och legala instanser.

Intervjuare (CD): **Access-isolering, vad menar du med det?**

Respondent: Det jag menar är att eftersom det är serverless så har du ju inget servicekonto direkt som styr detta, som man kan hantera via klassisk AD-säkerhet och liknande, det är det att vi måste förlita oss på att den här tjänsten inte blir hackad eller liknande.

Intervjuare (RF): **De begränsningarna du nämner, vad har dessa för effekt på organisationen och er?**

Respondent: På den generella organisationen har dem egentligen ingen effekt men på oss har de effekten att vi ständigt måste iterera delar av vår plattform för att vi kan ju inte bara slappna av och tänka att det kommer fungera i all framtid utan vi måste ständigt titta på vad skulle vi kunna ersätta detta med. Det är klart det ska man göra i vilket fall som helst bara det att detta här kryper närmre och närmre hela tiden för vi vill ju ha en lösning som är så smidig som möjligt dock med viss maintainance, det kommer man inte undan, men här är det mer att det finns risk att vi kommer lappa och lappa och lappa och till slut smäller alltihop, det är den största grejen och en rätt stor grej för att vara ärlig.

Intervjuare (CD): **Det du nämner kring att hitta alternativ till de verktyg ni använder idag, vad hindrar er organisation från att göra just detta?**

Respondent: Det som hindrar är att det faktiskt fungerar, det är väldigt låg kostnad. För att vi faktiskt ska titta på och hitta ett nytt verktyg så måste ROI:n för detta finnas och det är ju väldigt svårt det här med ROI, speciellt med tanke på att det inte är säkert att ROI:n är monetär utan det är mer att vi kommer att spara massa tid men det är inte alltid så att dem som betalar

håller med. Det är den typen av problem, hade vi bara haft så mycket budget som helst hade vi tagit en annan väg.

Intervjuare (RF): **Du pratar om att ni lappar ihop flera gånger med risk att det kan krascha, vilket, från vårt perspektiv uppfattas som en valid anledning att kolla på och implementera en annan lösning utifrån ett ledningsperspektiv.**

Respondent: Man kan tycka det! Men det är inte bara så att vi kan byta utan det är en lång process, man måste titta på och sätta upp ett koncept, man måste iterera igenom, man kan inte bara titta på en utan ett antal. Finns det inte en tjänst inom AWS ex. Så måste du prata med ett antal olika vendors, man måste gå igenom kapabiliteter, man måste mappa, man måste få fram avtal med dem för att man då kan göra en *proof of concept*, det är en lång process och det är många steg. Det låter enkelt men det är inte det. Kostnaden är ju mer än bara själva verktyget utan det är också allt runt omkring.

Intervjuare (CD): **Hur kompatibla är de verktyg som ni använder nu tillsammans med andra verktyg som ni potentiellt kan använda eller använder?**

Respondent: Så länge ett annat verktyg har connectors till Amazon S3, i detta fall, så är det inga problem eftersom all validering sker i data laken som i sin tur är lagrad på Amazon S3. Det är egentligen det enda kravet det är, sen såklart, om det är en icke-servless tjänst som ska köras så måste vi titta på det här med accessisolering osv.

Intervjuare (CD): **Anser du att det är främst tekniska eller organisatoriska problem ni upplever?**

Respondent: Jag anser att de största problemen är organisatoriska eller operationella för att faktiskt hantera informationen i sig är inga problem, det är inga problem att skyffla data, det är inga problem att förbereda data, det är inga problem att rengöra data, det finns en uppsjö olika verktyg det är egentligen bara hur du vill ha det, vill du att det ska vara snygg grafiskt eller att det ska vara terminalbaserat, det spelar ingen roll, oftast är problemet operationellt och även, framförallt, hur man klassificerar informationen, det är det som oftast är det stora problemet, att man garanterar att det inte kommer med känsligt information eller att man garanterar att rätt person eller rätt funktion får tillgång till informationen och rätt information och bara den informationen dem ska ha, det är det som är de största problemen. Skära i informationen och garantera att den kommer fram, det är inget konstigt, det är bara att bestämma hur ska vi skära, vad ska vi skära på, vad är känsligt, vad är inte känsligt, det är alla dem processerna som tar mest tid. Teknologin har funnits i många år det är bara det att det går snabbare framåt nu, det kommer mer kapabiliteter, det kommer fler som gör samma sak, man är inte längre tvingad till att använda dem få som fanns tidigare.

Intervjuare (CD): **Skulle du då uppskatta att ni har en större mängd överflödiga ostrukturerade data som i princip bara är en kostnad för företaget eftersom den inte kan hanteras på rätt sätt?**

Respondent: Både och för att det är så att vi har en massa data som vi inte tar itu med nu men där är ju planer för det för att informationen har blivit identifierad som potentiellt värdefull för en viss typ av uppföljning, analys, prediktioner. Den ligger och väntar men den är ändå organiserad, den är taggad, den är parsead, vi har det strukturella schemat på plats även om den är ostrukturerad så har vi metadata om den. Den är sökbar i vår Data Catalog som också är extremt viktig när man pratar Big Data, man måste ha en Data Catalog sen hur bred den är eller hur komplett den är spelar ingen roll, har man en Data Catalog som är sökbar och uppdateringsbar så har du kommit en väldigt lång bit på vägen gällande Governance. Det var något vi tidigt identifierade att utan det så kommer detta att krackelera och det hade det gjort, då hade det blivit ett Data Swamp.

[38:32] **Framtiden**

Intervjuare (RF): **Vilka är de kortsiktiga utmaningarna för er Big Data-hantering där du befinner dig nu?**

Respondent: Det är att lösa presentationen av informationen för att olika delar av verksamheter anser olika saker. Self-service är ett ord som betyder så olika grejer för olika personer, får du ett mejl med Excel som innehåller rådata, det är self-service för någon, ge mig direkt access till data laken och jag får använda vilka tjänster jag vill, det är self-service för någon annan. Problemet för oss är att hitta en lämplig nivå på vad vi ska presentera och hur vi ska presentera det eller göra det tillgängligt, detta är både kortsiktigt och långsiktigt eftersom det är något som inte försvinner, folk kommer alltid att vilja titta på information eller få tillgång till information på olika sätt och när vi väl ger dem tillgång till informationen på det sättet så kommer dem att börja transformera informationen för att passa deras behov och när dem väl börjar göra det så tappar vi lineage på datan. Vi har ett långsiktigt mål också att utöka vår data lineage, det är därför vi har tittat på Collibra bl.a. för att den då ska kunna scanna hela plattformen för att vi ska se och kunna hitta vad har hänt med datan på vägen för att vi ska kunna garantera kvalitén på datan och även kunna se att, okej, business säger att den här datan ni har gett till oss är dålig, det funkar inte vi får inte rätt grejer, då ska vi kunna peka på och se att, okej, men ni har transformerat eller filtrerat bort detta och detta här innan ni gör slutgiltig rapport, det är därför det inte stämmer, ni har förändrat kärnan av informationen, därför stämmer det inte eller rättare sagt, det är därför ni får dessa siffrorna och inte de andra siffrorna.

Intervjuare (CD): **Hur tror du att de begränsningar du nämnt innan kan påverka er i det långa loppet?**

Respondent: Det är lite som jag sa, jag tror att begränsningarna kommer påverka oss på det sättet att vi blir tvingade till att förr eller senare göra ett omtag på de här bitarna, antingen det eller så kommer Amazon att släppa efter på vissa av dessa saker som jag nämnt. Vi har haft diskussioner med dem också och tanken med lambda från början var att det skulle vara korta snabba funktioner som gör en viss uppgift medan vi inte använder lambda som Amazon tänkte från början men det betyder inte att det är felaktigt. Vi hoppas på att Amazon ska släppa på eller erbjuda en alternativ möjlighet, vi vill gärna hålla det serverless så långt det går, så lite management som möjligt när det gäller de här bitarna, vi vill kunna konfigurera dem, vi vill kunna monitorera dem, men vi vill inte behöva skruva på en massa grejer bara för att få det att fungera.

Intervjuare (RF): **Vi börjar känna oss hyfsat nöjda, är det något du vill tillägga som du tycker att vi missat att ta upp?**

Respondent: Nej, det skulle väl i så fall vara Data Lifecycle, management och sådana saker men det har vi ju varit inne på lite grann, data retention är också en viktig bit. Man kanske måste sätta upp sådana grejer också, vi har data retention uppsatt baserat på vad det är för typ av klassificering på informationen, till exempel information som är känslig har vi satt en fixad tidsram på som det får lov att finnas tillgänglig, medan information som är viktig ur ett legalt perspektiv inte har någon retention osv. Det är något som spelar roll.

[44:40] **Avslutning**

Intervjuare (CD): **Då signalerar jag att inspelningen avslutas.**

Referenser

Abawajy, J. (2015). Comprehensive analysis of big data variety landscape. International journal of parallel, emergent and distributed systems, 30(1), 5-14.

Adiba, M., Castrejon-Castillo, J. C., Oviedo, J. A. E., Vargas-Solar, G., & Zechinelli-Martini, J. L. (2016). Big data management challenges, approaches, tools and their limitations.

Akerkar, R. (Ed.). (2013). Big data computing. Crc Press [Hämtad 24 februari 2020]

Al Nuaimi, E., Al Neyadi, H., Mohamed, N., & Al-Jaroodi, J. (2015). Applications of big data to smart cities [Hämtad 26 mars 2020]

Almeida, F. L. (2017). Benefits, challenges and tools of big data management. Journal of Systems Integration, 8(4), 12-20 [Hämtad 28 mars 2020]

Apache Hadoop. (n.d.-a). The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing. Tillgänglig via: Apache https://hadoop.apache.org [Hämtad 25 mars 2020]

Apache Hadoop. (n.d.-b). MapReduce Tutorial. Tillgänglig via: Apache https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html [Hämtad 25 mars 2020]

Apache Kafka (n.d). Apache Kafka® is a distributed streaming platform. What exactly does that mean? Tillgänglig via: Apache https://kafka.apache.org/intro [Hämtad 25 mars 2020]

Barnaghi, P., Sheth, A., & Henson, C. (2013). From data to actionable knowledge: Big data challenges in the web of things. IEEE Intelligent Systems, 28(6), 6-11 [Hämtad 25 mars 2020]

Bottcher, E. (2018). What I Talk About When I Talk About Platforms. Tillgänglig via: Martin Fowler https://martinfowler.com/articles/talk-about-platforms.html [Hämtad 1 april 2020]

Bryman, A. (2016). Social research methods (Fifth edition ed.): Oxford University Press [Hämtad 2 maj 2020]

Cambridge Dictionary (n. d.). Meaning of Database in english. Tillgänglig via: Cambridge https://dictionary.cambridge.org/dictionary/english/database [Hämtad 1 april 2020]

Chand, S. (2019). What Is Talend? – An Unified Platform For Data Integration, Tillgänglig via: Edureka https://www.edureka.co/blog/what-is-talend-tool/#WhatIsTalend [Hämtad 25 mars 2020]

Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S., & Zhou, X. (2013). Big data challenge: a data management perspective. Frontiers of Computer Science, 7(2), 157-164 [Hämtad 18 februari 2020]

Computerworld. (2020). AWS vs Azure vs Google Cloud: What's the best cloud platform for enterprise? Tillgänglig via: ComputerWorld https://www.computerworld.com/article/3429365/aws-vs-azure-vs-google-whats-the-best-cloud-platform-for-enterprise.html [Hämtad 18 mars 2020]

DalleMule, L., & Davenport, T. H. (2017). What's your data strategy? Harvard Business Review, 95(3), 112-121 [Hämtad 3 mars 2020]

DataFlair (2019). Top 10 Big Data tools that you should know about. Tillgänglig: https://data-flair.training/blogs/top-big-data-tools/ [Hämtad 28 mars 2020]

Datamation. (2020). AWS vs. Azure vs. Google: Cloud Comparison. Tillgänglig via: Datamation https://www.datamation.com/cloud-computing/aws-vs-azure-vs-google-cloud-comparison.html [Hämtad 28 mars 2020]

Demchenko, Y., Grosso, P., De Laat, C., & Membrey, P. (2013). Addressing big data issues in scientific data infrastructure (pp. 48-55). IEEE [Hämtad 25 mars 2020]

Dictionary. (n.d.) Limitation. Tillgänglig via: Dictionary https://www.dictionary.com/browse/limitations [Hämtad 9 april 2020]
Dobre, C., & Xhafa, F. (2014). Intelligent services for big data science. Future generation computer systems, 37, 267-281 [Hämtad 22 februari 2020]
E. Murphy, R. Dingwall, D. Greatbatch, S. Parker, and P. Watson, "Qualitative Research Methods in Health Technology Assessment: A Review of the Literature, "Health Technol. Assess. (Rockv)., vol. 2, no.16, pp. 1–274, 1998 [Hämtad 22 februari 2020]
Edupristine (2017). Top and trending Hadoop tools in Big Data. Tillgänglig: https://www.edupristine.com/blog/top-big-data-hadoop-tools [Hämtad 28 mars 2020]
Emrouznejad, A. and SpringerLink (Online service) (2016) Big Data Optimization : Recent Developments and Challenges. Springer International Publishing (Studies in Big Data: 18). Tillgänglig via: LUBSearch https://search.ebscohost.com/login.aspx?direct=true&db=cat07147a&AN=lub.5615324&site=eds-live&scope=site [Hämtad 25 februari 2020]
Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. National science review, 1(2), 293-314 [Hämtad 3 april 2020]
Ferguson, M. (2012). Architecting a big data platform for analytics. A Whitepaper prepared for IBM, 30 [Hämtad 25 mars 2020]
Flick, U., von Kardoff, E., & Steinke, I. (Eds.). (2004). A companion to qualitative research. Sage [Hämtad 9 april 2020]
Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International journal of information management [Hämtad 24 februari 2020]
Gantz, J., & Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. IDC iView: IDC Analyze the future, 2007(2012), 1-16 [Hämtad 24 februari 2020]
Hu, F. (2016) Big Data: Storage, Sharing, and Security. CRC Press. Tillgänglig via: LUBSearch https://search.ebscohost.com/login.aspx?direct=true&db=cat02271a&AN=atoz.ebs10416891e&site=eds-live&scope=site [Hämtad 24 februari 2020]
Jacobsen, D. I. (2002). Vad, hur och varför : om metodval i företagsekonomi och andra samhällsvetenskapliga ämnen: Studentlitteratur [Hämtad 24 februari 2020]
Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. Proceedings of the VLDB Endowment, 5(12), 2032-2033 [Hämtad 25 mars 2020]
Lakoju, M. and Serrano, A. (2017) 'Saving costs with a big data strategy framework', 2017 IEEE International Conference on Big Data (Big Data), Big Data (Big Data), 2017 IEEE International Conference on, pp. 2340–2347 [Hämtad 22 februari 2020]
M. Myers, Qualitative research in business and management. Sage, 2009 [Hämtad 22 februari 2020]
McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. Harvard business review, 90(10), 60-68 [Hämtad 18 februari 2020]
MongoDB (n.d.-a). The database for modern applications. TIllgänglig via: MongoDB https://www.mongodb.com/ [Hämtad 26 mars 2020]
MongoDB (n.d.-b). Introduction. Tillgänglig via: MongoDB https://docs.mongodb.com/manual/introduction/ [Hämtad 26 mars 2020]
Oates, B.J. (2005). Researching information systems and computing. London: Sage Publications Inc. [Hämtad 22 april 2020]
Oracle. (n.d.-a). What is Big Data? Tillgänglig via: Oracle https://www.oracle.com/big-data/guide/what-is-big-data.html#link3 [Hämtad 17 mars 2020]

Oracle. (n.d.-b). Oracle Data Mining: Scalable in-database predictive analytics. Tillgänglig via: Oracle https://www.oracle.com/database/technologies/advanced-analytics/odm.html [Hämtad 27 mars 2020]

Raheem, N. (2019). Big Data. New York: Chapman and Hall/CRC. Tillgänglig via: LUBSearch https://doi-org.ludwig.lub.lu.se/10.1201/9780429060939 [Hämtad 3 mars 2020]

Rubin, V., & Lukoianova, T. (2013). Veracity roadmap: Is big data objective, truthful and credible?. Advances in Classification Research Online, 24(1), 4 [Hämtad 17 mar 2020]

Russom, P. (2011). Big data analytics. TDWI best practices report, fourth quarter, 19(4), 1-34 [Hämtad 3 mars 2020]

Russom, P. (2013). Managing big data. TDWI Best Practices Report, TDWI Research, 1-40 [Hämtad 17 mars 2020]

SINTEF. (2013, May 22). Big Data, for better or worse: 90% of the world's data generated over the last two years. ScienceDaily. Tillgänglig via: Google Scholar www.sciencedaily.com/releases/2013/05/130522085217.htm [Hämtad 17 mar 2020]

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. Journal of Business Research, 70, 263-286 [Hämtad 18 februari 2020]

Tableau. (n.d.). What is Tableau? Tillgänglig via: Tableau https://www.tableau.com/sv-se/products/what-is-tableau [Hämtad 28 mars 2020]

Talend. (n.d.-a). What is ETL (Extract, Transform, Load)? Tillgänglig via: Talend https://www.talend.com/resources/what-is-etl/ [Hämtad 31 mars 2020]

Talend. (n.d.-b). What is Data Processing? Tillgänglig via: Talend https://www.talend.com/resources/what-is-data-processing/ [Hämtad 1 april 2020]

Vaghela, Y. (2018). Four Common Big Data Challenges. Dataversity. Tillgänglig via: Dataversity https://www.dataversity.net/four-common-big-data-challenges/ [Hämtad 27 mars 2020]

Yi, X., Liu, F., Liu, J., & Jin, H. (2014). Building a network highway for big data: architecture and challenges. Ieee Network, 28(4), 5-13 [Hämtad 22 februari 2020]

Zhang, F., Liu, M., Gui, F., Shen, W., Shami, A., & Ma, Y. (2015). A distributed frequent itemset mining algorithm using Spark for Big Data analytics. Cluster Computing, 18(4), 1493-1501 [Hämtad 18 februari 2020]

Zicari, R. V. (2014). Big data: Challenges and opportunities. Big data computing, 564, 103 [Hämtad 24 februari 2020]