En introduktion till Big Data

VISMA

Sammandrag

Big data har både blivit ett buzzword och namnet på ett fenomen som sannolikt kommer att öka i betydelse. I allt högre grad genereras och tillgängliggörs enorma mängder data som kan användas både isolerat och i kombination som styrning- och beslutsfattande verktyg för organisationer.

Det är inte nödvändigtvis enkelt för en organisation att använda big data: Det måste finnas ett strategiskt syfte bakom beslutet att använda informationen, och organisationen måste ha rätt kombination av talang, vilja kultur och teknik.

Big data innebär både risk och vinst. Naturligtvis är det tekniska och praktiska frågor som måste besvaras, men känsliga och komplicerade aspekter som säkerhet, integritet och mänskliga fel kan också leda till oförutsedda och komplicerade utmaningar. Dessutom bör man ha en viss förståelse för statistik. Å andra sidan, vinsten med att använda big data på rätt sätt är enormt gynnsam - och förväntas vara en nödvändighet för att vara konkurrenskraftig på lång sikt.

Den stora mängden data som är i omlopp har ett stort värde - men för vem? Det är inte omöjligt att den personliga informationen som individen generar på sikt kommer att vara en omsättningsbar tillgång på en börs av något slag. Varje Facebookanvändare är exempelvis i genomsnitt värd cirka 830 SEK - vilket Facebooks aktieägare idag kan glädja sig åt, men i framtiden är det kanske möjligt att den enskilda användaren också får en bit av kakan?

Big data väcker också en del etiska frågor. När man plötsligt har möjlighet att förutsäga framtiden med en viss sannolikhet, hur kommer det att påverka lagstiftningen? Etisk och moraliska dilemman får en ny dimension. De många användningsområdena som big data erbjuder begränsas endast av fantasin. “The Internet of Things”, datatrafik, sociala medier, sök- och surfhistorik, köpbeteende, personlig ekonomi, finansiell information, offentlig statistik, väderdata, navigationsdata och så vidare - möjligheterna att ta reda på saker är oändliga.

För att förverkliga allt detta behöver vi teknik. Kraven för hårdvara, databaser, mjukvara och relaterad kompetens är många och höga, och tekniken förändras snabbt

Om man tittar på marknaden och leverantörer av big data-teknologi ser man att amerikanerna är långt före världen i övrigt. De stora IT-jättarna har sett att big data är ett snabbt växande segment inom IT och har kastat sig in i kampen om marknadsandelar mot de små, nya nischleverantörerna. Det enda som är säkert, är att vi kommer att ha en spännande och utmanande tid framför oss!

Introduktion

«Big data» som vetenskapsgren är relativt ny och kunskapen är relativt begränsad. Syftet med detta dokument är att ge en allmän introduktion till vad “big data” är och förhoppningsvis inspirera till att vilja fördjupa sig mer i ämnet vilket garanterat kommer att vara ett relevant område under många år framöver.

Dokumentet tar upp definitionen av big data, risk och avkastning, exempel på praktiskt användande, introduktion till teknik och lösningar som möjliggör big data.

Innehållet i detta dokument är på en översiktlig nivå, vilket innebär att läsaren nödvändigtvis inte kommer att få en djup insikt i specifika ämnen inom big data. Däremot bör läsaren få en bättre förståelse för vad big data är, vilka värden och fallgropar som finns, samt en tydligare bild av vilka lösningar som finns och vem som levererar dem.

Big data är ett område i kontinuerlig förändring. Politiska och sociala krafter, som har mycket att säga till om hur big data används, kan ändra sig - den data som är lagligt och acceptabelt att samla in idag kanske inte är samma imorgon. Företag förändras, går i konkurs och nya uppstår - de som dominerar branschen idag kan befinna sig i en annan position imorgon.

Trevlig läsning!

Vad är Big Data?

“You can’t manage what you don’t measure” är ett känt uttryck från bland annat ledningsteorier; utan data kan man alltså inte styra. Med big data talas det om stora mängder kvantitativ data - och data är i grund och botten ett styrning- och beslutsstödsvertyg.

Big data representerar en revolution inom kultur, beslutsprocesser och strategi - inte bara teknik. (1)

Under hundratals år har ett antal individer, organisationer och institutioner samlat in väderdata. Tanken bakom datainsamlingen är bland annat eftersom vädret följer cykler kan man förutse vädret i framtiden genom att titta på historiska data. Samma sak gäller big data: genom att titta på historisk data får man en bättre grund att förutsäga framtiden. Med ökande bredd och djup av data kommer också - i teorin - större möjligheter att analysera sig fram till bättre beslutsfattande.

Även om många pratar om “big data”, betyder egentligen big data datauppsättningar som är så stora och komplexa att de är svåra att hantera med konventionella verktyg.

Gartners definition

Beroende på vem du frågar kommer du få olika definitioner av vad big data är. Det enklaste är att förhålla sig till Gartners definition med “3V” (2):

- Stor volum (volume).
- Stor variation av information (variation).
- Snabb förändring (velocity).

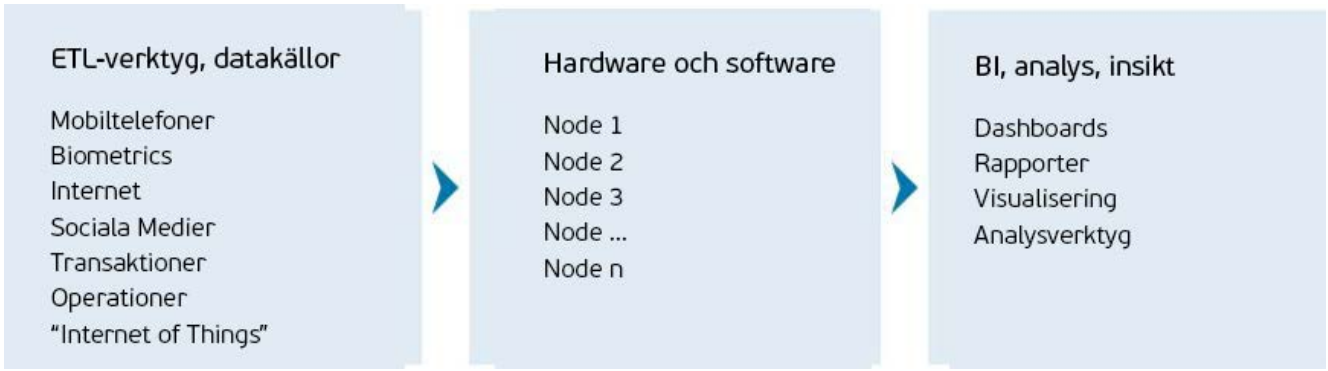
Big data är alltså stora och varierande mängder data som ständigt förändras.

Generering av data

Data måste hämtas in på ett eller annat sätt. ETL-enheter (extract, transform, load) hämtar data och matar in det i databaser, antingen direkt eller via andra databaser. Internetåtkomst är en kritisk komponent för denna datagenerering. När allt fler objekt blir anslutna till internet kommer en ökad mängd data att genereras kontinuerligt. Exempel inkluderar allt från GPS-/pulsklockor och andra biometriska mätare, mobiltelefoner, väderobservationer och andra typer av sensorer till sociala medier, CRM-system och internetbanker. Kort sagt - allt som genererar data.

Nästa steg i processen är att samla all data i en eller flera databaser, så kallad "data marts" eller datalager. Dessa lagringsmedier kan befinna sig på ett eller flera ställen, och fungera både on-premise och i molnet.

I den andra änden av värdekedjan behöver man verktyg för att skapa business intelligence, analys och insikt genom användning av dashboards, rapporter, visualisering av data och andra analysverktyg.



1. "2013-02 Academic Keynote Panel - THE REALITY OF BIG DATA", MIT Sloan CIO Symposium Videos, 15 August 2013
2. http://en.wikipedia.org/wiki/Big_data

Vanlig data vs big data

Skillnaden mellan "vanlig data" och "big data" beror i praktiken på organisation och person. För små företag kan till med några få gigabyte data vara “big”, medan Facebook eller Google pratar om petabyte och miljarder rader med data när de refererar till “big data”.

Förutom storlek, kan data också delas in i två huvudkategorier:

- Strukturerad
- Ostrukturerad

Medan strukturerad data är det man vanligtvis hittar organiserat på ett fördefinierat sätt i en databas, är ostrukturerad data "allt möjligt" som till exempel bilder, ljudfiler, filmer, e-post, dokument, kalkylark och annan data som inte är organiserad.

Många menar att andelen ostrukturerad data är cirka 70-90% av all data. Data-mining, textanalys, taggning och annan användning av metadata är exempel på hur man strukturerar ostrukturerad data.

Dagens BI-lösningar kräver i hög grad strukturerad data; big data är ofta ostrukturerad och kommer att kräva andra lösningar för att kunna konverteras till värdefull information. Medan BI-lösningar tenderar att erbjuda beskrivande statistik, erbjuder big data explorativ och kausal statistik.

Det är emellertid inte datamängden eller typen av data som är big datas stora värde, utan snarare möjligheterna att aggregera och bearbeta enorma datamängder. Detta beror på möjligheterna som uppkopplade system i kombination med ständigt ökande processorkraft ger.

Med big data kan mängden data nästan vara obegränsad, så om man letar efter nålen kan det vara klokt du vill göra höstacken lite mindre - dvs. begränsa dig själv (3). Oavsett omfattning och typ, handlar big data om att fånga, bearbeta, lagra, söka, dela, överföra, analysera och presentera (populärt kallat "visualisera") information. För att lyckas med detta måste man bland annat göra utvärderingar med hänsyn till underliggande teknik, kompetensen och de resurser som finns tillgängliga, och inte minst vilken typ av insikt du vill ha.

Insikt - eller undersökningsdesign, för att vara mer specifik - kan delas in i tre kategorier (4):

- Explorativ - eller utforskande: försök att förstå, tolka och genererahypoteser. Bäst lämpad för kvalitativ datainsamling, till exempel fokusgrupper.

- Deskriptiv - eller beskrivande: test av hypoteser och upprätta korrelationer (men inte sammanhang!). Lämplig för kvantitativa data, särskilt big data.

- Kausalitet, eller orsakssamband: en form av experiment där man fastställer att en händelse (x) är orsaken till en annan händelse (y) givet en uppsättning villkor (z). Big data passar inte särskilt bra för den här typen av insikt.

Big data i organisationer

Big data är inte något du bara börjar använda; det bör drivas av en väl tänkt process om du vill uppnå varaktiga, positiva förändringar. Två viktiga variabler inkluderar att organisationen har en medveten relation till varför de vill använda big data och att ledningen måste vara medveten om de utmaningar de står inför om de väljer att gå den vägen.

Fyra frågor som ledningen bör ställa sig:

Fråga 1:
Vad är det vi kan vi dra nytta av att veta som data kan berätta?

1. Bättre beslutsunderlag, eftersom man kan gå från "vad tror vi" till "vad vet vi" och därmed förändras från att styras av antaganden till att vara "datadriven".

2. Upptäck nya (dolda) insikter, eftersom man kan börja hitta relationer som hittills inte varit möjliga att upptäcka.

3. Automation av affärsprocesser. Stora produktionsföretag kan med hjälp av big data få en tydligare och mer sammanhängande överblick över värdekedjan, från råvaror som kommer in, genom produktionsprocessen och till logistik för att få den färdiga produkten vidare till kunden. Var uppstår flaskhalsarna och hur optimerar man processerna?

Fråga 2:
Var finns datan?

Databasen som behövs för att bygga den insikt man vill ha måste vara tillgänglig. De flesta organisationer har mycket data i sina egna system (både strukturerad som i försäljnings- och ekonomisystem) och genom externa källor (SCB, finansiella siffror, kreditvärderingsinstitut, sociala medier etc.).

Fråga 3:
Hur ska data/information produceras?

Data måste samlas in, aggregeras, bearbetas och rapporteras till olika intressenter - både internt och kanske även externt. Här är det många överväganden att ta hänsyn till, både med tanke på teknologi, lösningar och expertis och användningsområde. I många fall kommer organisationen att sakna den nödvändiga kompetensen internt, och därmed vara tvungen att göra en bedömning om de ska bygga upp kompetensen internt eller ta in konsulter (“make or buy”).

Fråga 4:
Hur ska resultatet från datainsamlingen kommuniceras?

I värsta fall skapas rapporter som blir liggandes i mailen och glöms bort, medan det i bästa fall skickas ut insikter från big data till de anställda som har möjlighet att agera utifrån det (till exempel en säljare som automatiskt ser en dashboard med KPI:er till en potentiell eller befintlig kund.) Det handlar lika mycket om att ge högsta ledningen insikter som att ge organisationen möjlighet att vidta åtgärder baserat på fakta. På så sätt är det viktigt att göra bra information tillgänglig på ett så förståeligt sätt som möjligt för mottagaren. Visualisering av data benämns ofta som ett relativt enkelt sätt att säkerställa att fler förstår vad de ser. Man ska berätta historier med visualisering. Hans Rosling, en av de skickligaste på att förmedla komplicerad, statistikbaserade budskap på ett sätt som är lätt att förstå för de flesta (5).

Konsulterna inom Visma bWise kan hjälpa till med alla dessa frågor.

3. McGuire, T., Meyer, C. & Stone, D. (2014), “The data-driven life”, McKinsey on Marketing & Sales, Oktober 2013
4. Gripsrud, G. Olsson, U.H., & Silkoset, R. (2010), “Metode og dataanalyse”, Høyskoleforlaget
5. Dokumentarvideo: “Vitenskapens verden: Hans Rosling om befolkningseksplosjonen” (vist på NRK2 30. mars 2014)

Fyra utmaningar som ledningen bör vara medveten om

Utmaning 1: Talang

Kompetensen om big data måste utvecklas. Särskilt de som är bra på att visualisera data och de som kan se sambandet mellan data och affärsmöjligheter är av högt värde. Att kunna bearbeta, lagra och distribuera data på ett så tryggt och säkert sätt som möjligt.

Utmaning 2: Vilja

Ledningen måste vilja använda big data.

Utmaning 3: Organisationskultur

För att lyckas med big data finns det mycket som beror på organisationens vilja och dess personer (6), såväl som beslutsprocesser och kultur (7) - tänker organisationen "vad tror jag" eller "vad vet vi"? Accenture räknar med att över hälften av alla statliga ledare saknar grundläggande förståelse för analys och tillämpning av insikter (8). Detta är en betydande utmaning och det är allvarligt om omfattningen faktiskt stämmer.

Utmaning 4: Teknik

Ostrukturerad data och flera datakällor bjuder på utmaningar.

Implementering

En välkänd kliché är att majoriteten av IT-projekt inte går som planerat, varken när det gäller tidsramar, budget eller funktionalitet (9). Data och information kommer inte av sig självt - därför är det viktigt att alla inblandade är involverade från början av implementeringen. En del forskning, till exempel från MIT Sloan Business School (10) visar att datadrivna organisationer är betydligt mer produktiva och lönsamma än icke-datadrivna organisationer. Det borde vara bra motivation att lägga ner engagemang på ett bra genomförande.

Framgångsrik implementering av en big data-strategi kräver bra teknik i botten och bra människor som kan dra nytta av möjligheterna (antingen interna eller konsulter). Organisationer som har en bra och tydlig strategi, som har experter både på teknik och siffror kommer att lyckas extra bra.

Det fyra steg man måste gå igenom när man implementerar big data-lösningar:

1. Skapa en gemensam målbild av vad big data-implementering bör leda till
2. Prioritera vad big data-verktyg ska användas till
3. Hitta modeller och verktyg som stödjer målbilden och prioriteringarna
4. Använd lösningarna på ett sätt som är tillgängligt för alla i organisationen

6. Barton, D. & Court, D. (2012), "Making Advanced Analytics Work for You", HBR, October 2012
7. McAfee, A. & Brynjolfsson, E. (2012), "Big Data: The Management Revolution", HBR, October 2012
8. "Government Analytics: What Governments Stand to Gain (or Lose)" av Accenture (Christopher Gray), 2012
9. "McKinsey Report Highlights Failure of Large Projects", Gartner Blogs (Mark McDonald), October 2012
10. "Erik Brynjolfsson on Big Data", MIT Sloan Business School, February 2012

Värdet av Big Data:
Risk, avkastning och etik

Risk

När man överväger att investera i eller implementera stora datalösningar, finns det åtminstone fyra tekniska och praktiska aspekter att beakta:

- **Kostnader**
  Etablering, underhåll/drift/uppdatering, hårdvaruressurser, omställningskostnader, utbildning.

- **Teknik**
  Filformat, datatillgänglighet, expansionsmöjligheter (kvantitet och datatyp), anslutningsmöjligheter.

- **Säkerhet**
  Åtkomst, tillgänglighet, integritet.

- **Användarvänlighet**
  Intern kompetens/erfarenhet, oberoende expertis, automatisering, återanvändning.

Att arbeta med big data kan bli dyrare än väntat, det kan vara för tekniskt komplicerat att hantera, säkerhet kan visa sig vara svårt att underhålla och det kan vara för svårt att applicera på de i organisationen som kommer att arbeta med detta dagligen.

Vi måste tänka framåt både om de åtgärder vi kommer att ta och eventuella konsekvenser det kan få. Sannolikheten att något allvarligt och oväntat kommer att inträffa imorgon är kanske liten, men konsekvenserna är stora.

Förutom ovanstående steg finns det två typer av risker sin man kan stöta på (12): Säkerhet och integritet, och mänskliga misstag.

11. Taleb, N.N. (2012), "Antifragile: Things that Gain from Disorder", Amazon Kindle
12. "My data is bigger than yours. Analyzing the fragility of BigData with Nassim N. Taleb", Valbonne Consulting (Joachim Bauernberger), Oktober 2013

Säkerhet och integritet

Om alla litade på varandra, skulle säkerhet inte vara ett tema. Exempel på fusk och bedrägeri är oändligt många, men det finns också några intressanta exempel på hur människor kan vara ärliga:

"Bagelmannen" som vi kan läsa om i boken Freakonomics (13), levererade bagels till kontor i Washington varje morgon, med en korg för att lägga pengar i, och kunderna kunde betala så mycket eller så lite som de ville. Bland de många intressanta upplevelser han har fått, är den största att när människor är fria att betala vad de känner för (eventuellt inte betala), då betalar folk för sig - således de är ärliga.

Bandet Radiohead släppte år 2007 albumet "In Rainbows" där de som laddade ner albumet kunde betala så mycket eller så lite som de ville (14): De flesta betalade inte för albumet, men fortfarande under de tre första månaderna sålde albumet mer än Radioheads tidigare album ("Hail To The Thief") gjorde totalt. (Varken Radiohead eller andra grupper har därefter valt samma affärsmodell).

Om vi alla kunde lita på varandra, skulle det inte finnas en marknad för säkerhet - men det kan vi uppenbarligen inte. Med andra ord, misstro har skapat en marknad för säkerhetstjänster (15).

Dataskyddsmyndigheten definierar tre huvudområden när det gäller informationssäkerhet och risk, som till stor del inkluderar personlig integritet: sekretess, integritet och tillgänglighet.

Särskilt efter Snowden-avslöjandet har säkerhet blivit en av de mest aktuella ämnena, både i professionell fackpress och i medier i övrigt. Som chef för den norska datatillsynen, Bjørn Erik Thon, konstaterar i Dagens Næringsliv 15 juli 2014 (16): De flesta har blivit mycket mer medvetna om sin personliga integritet efter Snowden-avslöjandet, medan vi inte nödvändigtvis förstår vilken typ av data vi tillhandahåller om oss själva och vi kan omöjligt förutsäga konsekvenserna av kränkningar av den personliga integriteten (17).

Företag som jobbar med big data skulle gynnas av att vara både vara ärliga och tydligt kommunicera till sina kunder att de tar integritet på allvar. Alternativet är negativa spridningar i media, som Facebook fick en del av sommaren 2014 när det blev känt att de samlade in mer användarinformation än man först trodde (18).

Mänskliga fel

Oavsett nivå på system, verktyg och rutiner, när det gäller säkerhet, kan man inte helgardera sig mot mänskliga fel. Har du låst bilen och lämnat nycklarna på taket, skrivit PIN-koden på ditt kreditkort, glömt en utskrift med konfidentiell information, glömt kvar datorn på flygplatsen eller på något annat sätt öppnat dörren för att information ska komma på villovägar så är den potentiella skadan redan gjord. Det läskiga med big data är att du kanske inte förstår vilken data som kan komma på avvägar och vad det kan betyda.

13. Dubner, S.J. & Levitt, S. (2004), "What the bagel man saw", The New York Times
14. "Radiohead reveal how successful 'In Rainbows' download really was", NME, 15. oktober 2008
15. Schneier, B. (2012), «Liars and Outliers: Enabling the Trust that Society Needs to Thrive", Amazon Kindle
16. "Personvern etter Snowden", Dagens Næringsliv, Bjørn Erik Thon, 15. juli 2014
17. "When data gets creepy: The secrets we don't realise we're giving away", The Guardian, 5. desember 2014.
18. "Slik kartlegger Facebook meldingene dine", Dagens Næringsliv, 31. juli 2014

Okunnighet hos individen

Det är svårt att förstå konsekvenserna som följer av att all slags data och information kan lagras, indexeras och vara sökbar för alltid.

Något så enkelt som att använda en träningsapp med GPS-spårning, till exempel Strava eller Garmin Connect, kan avslöja för tjuvar var värdefulla cyklar kan stjälas (19). Genom att använda en app som Creepy kan (eller kunde) man till exempel söka efter ett Twitter-användarnamn och se användarens foton och geografiska platser - allt baserat på offentlig information (20). Ett antal liknande tjänster finns tillgängliga online, och belyser hur glidande övergången från "jag har inget att dölja" till att man lämnar ut obekvämt mycket information.

En utmaning är att viss data är tillgänglig; det är något helt annat om aggregerad data - den insikt du kan få om individen kan snabbt bli mycket mer än man önskar. Om en okänd person samlade in alla dina e-postmeddelanden, kontaktlista, Facebookprofil, kreditkortshistorik, besökta platser och dina fotoalbum - vad kommer denna person att få veta om dig? Föreställ dig att det fanns en app som tillät dig att se vem som är omkring dig och som samlade in och presenterade all tillgänglig information om användaren? Appen "Singles Around Me" (21) är ett kusligt exempel på vad man kan göra med informationen som användaren antagligen omedvetet lämnar efter sig.

Statistiska fel av analytikern

Mänskliga fel inträffar också hos den person som ska behandla uppgifterna.

Fooled by randomness (22): det är möjligt att få samma statistiska fel med stora datauppsättningar som med små datauppsättningar. De är inte representativa och man gör upptäckter som inte nödvändigtvis matchar verkligheten. Dessutom, stora datauppsättningar kan innehålla väldigt mycket brus i förhållande till inflöde av data, vilket innebär att det man faktiskt borde följa upp drunknar i allt brus.

Big data gör det enkelt att hitta relationer mellan variabler som egentligen inte har med varandra att göra; även om antalet skogsbränder och glassförsäljningen ökar på samma gång på sommaren, betyder det inte att folk vill köpa glass eftersom skogen brinner eller att glass triggar pyromani - orsaken är bara att det är varmare. Detta är något folk kan förstå, men inte datorer. Business Insider hade en intressant video som understryker att korrelation inte nödvändigtvis betyder att det finns ett sammanhang: till exempel pendeldistansen till jobbet och andelen företag som ägs av kvinnor (r = 0,71) och fans av artisten Prince har fetare barn(r = 0,77) (23).

1936 förutspådde The Literary Digest att republikanska presidentkandidaten Landon skulle vinna presidentvalet mot demokraternas Roosevelt, baserat på en undersökning med 2,3 miljoner svarande, genomförd via telefon (24). Men Roosevelt vann överlägset, och felet med undersökningen var att de som svarade kontaktades via telefon - en teknologi endast för de rika, som i stor grad sympatiserar med republikanerna. Liknande urvalsfel kan uppstå idag om man kritiskt hämtar sina svarande från till exempel Twitter eller Facebook.

19. "Advarer syklister mot treningsapper med GPS", VG, 16. mars 2014
20. Applikasjonen "Creepy" er tilgjengelig hos CNET
21. Applikasjonen "Singles Around Me" er tilgjengelig i Google Play
22. "Fooled by Randomness" av Nassim Taleb (http://www.amazon.com/dp/B001FA0W5W)
23. "6 Compelling Correlations That Make Absolutely No Sense", Business Insider, 5. desember 2014.
24. "Opinion poll", Wikipedia

Return (avkastning)

Värdet - där avkastningen mäts - definieras på många sätt, men på ett ekonomiskt eller finansiellt perspektiv betyder värdet nyttan en produkt eller tjänst ger. För aktörer inom näringslivet och särskilt i den privata sektorn mäts värdet främst i kronor och öre som tillfaller aktieägarna och ägare. I den offentliga sektorn kan värdet avse mer effektiv användning av offentliga medel. Värdet för slutkunder och konsumenter uttryckt i pengar, tid och funktionella och symboliska variabler.

Big data skapar värde på det sätt att det kan ge bättre beslutsunderlag, avslöja dolda insikter och eller förbättrade affärsprocesser (som nämnts tidigare), vilket i sin tur kan bidra till ökat värdeskapande. Tre element måste vara på plats för att datan ska vara värdefull för beslutsfattare (25):

- Datans användningsområde måste vara tydlig: detta är kanske den största utmaningen idag - men när organisationer vänjer sig vid att hantera stora mängder data kommer det också att ge dem en bättre förståelse för vad de ska göra gör med det.

- Uppgifterna måste finnas: de flesta organisationer har tillgång till mycket data, både internt och från externa källor.

- Uppgifterna måste bearbetas: data i råformat har liten nytta i sig själv om det inte bearbetas för att tjäna ett syfte.

Den kritiska faktorn för värdeskapande i big data är kompetens. Människor med rätt kompetens är det dock brist på. Man behöver kvalificerade människor med bakgrund och nyfikenhet som behövs för att göra upptäckter genom att arbeta med big data (dvs att kunna koda och vara nyfiken). Kvalifikationer kommer inte nödvändigtvis från utbildningsinstitutioner. "Data scientists" är bekväma med ostrukturerad data.

Efterfrågan på personer som har expertis i hur man ska bearbeta datan är betydande (26), något man bland annat ser på omsättningen till de som levererar produkter och tjänster (27): Från 2012 till 2013 ökade marknaden totalt från 12 till 19 miljarder USD (61% tillväxt), varav hårdvaruintäkter ökade omsättningen från cirka 4 till 7 miljarder USD (65% tillväxt), programvara ökade från cirka 2 till 4 miljarder USD (85% tillväxt) och tjänster från cirka 5 till 7 USD (46%). Organisationer runt om i världen spenderar mycket pengar - och kommer att spendera mer på big data-vågen, och det är kanske inte så konstigt konsultföretag som McKinsey betonar att det är viktigt för organisationer att börja använda big data så snart som möjligt tills det inte längre ger någon jämförande fördel.

På kort sikt kommer därför värdet som big data skapar fångas av dem som levererar teknik och expertis, men på längre sikt - när kunskapsnivån och användningsområdet blir tydligare och mer integrerade i organisationerna - kommer värdet ligga i själva datan.

25. Cukier, K. & Mayer-Schonberg, V. (2012), “Big Data”: A Revolution that Will Transform How We Live, Work and Think”, Amazon Kindle

Personlig information = omsättbara tillgångar?

Det finns mycket pengar i big data och information, men information kommer att vara något individen äger och omsätter, eller kommer det att fortsätta vara något som leverantörerna äger och tjäna pengar på?

Ett intressant sätt att se vilket värde information (eller data) har är att jämföra hur mycket ett big data-företag prissätts på börsen (marknadsvärde) mot företagets bokförda värde. Ju högre talet är, desto mer av företagets värden ligger i icke-materiella tillgångar. Det kan vara förväntningar på ökade intäkter i framtiden eller i en tillgång som företaget har som inte är kvantifierat i bokfört värde, något som användardata till exempel kan vara.

Tre representativa big data-företag som är börsnoterade - och därmed har finansiell information tillgänglig - är LinkedIn, Twitter och Facebook. Värdet på dessa företag ligger i deras användare - det vill säga du och jag och alla andra registrerade, aktiva användare av deras tjänster.

Om vi jämför marknadsvärdet mot bokfört värde med LinkedIn, Twitter och Facebook, ser du att förhållandet är mellan 7x och 16x. Som jämförelse är det förhållandet mellan andra stora amerikanska företag långt under 5x och ofta nära 1x (med undantag för Tesla, med ett marknadsvärde på otroliga 42x).

LinkedIn, Twitter och Facebook har marknadsvärden på respektive 19, 23, 170 miljarder USD medan motsvarande bokförda värde är "bara" 3, 3, 17 och 10 miljarder USD (juli 2014). Sagt på ett annat sätt, 343 miljarder dollar (16, 20 respektive 153) är värden som marknaden är villig att betala för men som inte är bokfört.

26. Cukier, K. & Mayer-Schonberg, V. (2012), “Big Data”: A Revolution that Will Transform How We Live, Work and Think”, Amazon Kindle
27. http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017
28. “Putting big data and advanced analytics to work”, McKinsey & Company, Oktober 2013

I svenska kronor motsvarar skillnaden mellan marknadsvärde och bokfört värde för dessa fyra företag till 2258 miljarder SEK - nästan dubbelt så stor som den norska statsbudgeten var 2014. Dessutom beaktas marknadsvärdet relativt antalet användare kommer man att se något intressant:

- En LinkedIn-användare är i snitt värd $64 (600 SEK)
- En Twitter-användare $113 (1070 SEK)
- En Facebook-användare $133 (1200 SEK)

Snapchat är en annan populär tjänst inom sociala medier som har ett enormt värde: i november 2013, erbjöd Google 4 miljarder US$(29), eller cirka 37 Miljarder SEK. Ungefär samtidigt spekulerades det att Snapchat hade mellan 30 och 60 miljoner aktiva användare (30). I så fall kommer värdet på en Snapchat-användare att vara i genomsnitt vara mellan 400 och 900 SEK, i linje med andra sociala medier.

En genomsnittlig användare av LinkedIn, Twitter, Facebook eller Snapchat är alltså värd några hundralappar. Som användare kan du få tillgång till olika tjänster "gratis" i utbyte mot att släppa personlig information. Denna information använder företagen för att tjäna pengar, till exempel genom att sälja annonsutrymmen: Ju mer användaren ger ifrån sig av sin information, desto mer riktad blir annonseringen. Integriteten säkras (?) idag genom att användaren godkänner ett användaravtal. Här står det naturligtvis ingenting om att tjänsteleverantören ger användaren en del av vinsten eller det värde som användaren skapar - värdet tillfaller företaget och aktieägarna helt och hållet.

Å andra sidan, om den genomsnittliga Facebook-användaren inte är villig att betala mer än 875 SEK för att använda tjänsterna, vill Facebook naturligtvis anse att dagens affärsmodell (som ofta hotar privatlivet) som mer gynnsam. Om den genomsnittliga användaren någon gång i framtiden är villig att betala ett engångsbelopp eller en årlig avgift som tillsammans uppgår till ett värde högre än 875 SEK (och därmed skyddar integriteten bättre än idag) kommer Facebook ompröva affärsmodellen. Exempel på tjänster där tjänsteleverantören balanserar annonsering med betalning från slutanvändaren är Spotify och Strava. Du kan välja att inte betala något för att använda tjänsterna, men kommer sedan vara tvungen att finna sig i mer (riktad) reklam, vara produkten (informationen) du säljer till andra och / eller finna sig i att få reducerad funktionalitet.

Man kan tänka sig att användare av denna typ av tjänst i framtiden kommer att betalas för att vara en del av en social plattform, där du får ersättning för var hur mycket information som du ger ut - och att man som användaren flyttar sin datauppsättning dit man får mest värde.

Författarna av "Big Data" (31) spekulerar i att man i framtiden, kan omsätta information om sig själv på en börs - det är en intressant tanke där ansvaret för integritet överförs i större utsträckning till slutanvändaren.

	TWITTER	LINKEDIN	FACEBOOK
# användare (miljoner)	200	300	1280
Market cap (USD mdr)	23	19	170
Book value (USD mdr)	3	3	17
Market-to-book	7,70	6,91	10,17
Market cap per användare (USD)	113	64	133
Book value per användare (USD)	15	9	13

(Baserat på siffror och valutakurser i jul-aug 2014 från Wikipedia och http://finance.google.com)

29. Wikipedias side om Snapchat
30. "The truth about Snapchat's users", Business Insider, December 2013
31. Cukier, K. & Mayer-Schonberg, V. (2012), "Big Data": A Revolution that Will Transform How We Live, Work and Think", Amazon Kindle

Etik

Det ganska kända exemplet på den amerikanska "Target girl" som fick erbjudanden för mamma- och babyprodukter eftersom hennes köpmönster matchade profilen för gravida kvinnor (32), detta belyser möjliga etiska dilemma som ett resultat av möjligheterna för prediktiv analys som big data erbjuder. Om vi antar en databas för ett stort urval av världens befolkning (Facebook har över en miljarder användare, eller 15 procent av världens befolkning).

I vilken utsträckning ska vi kunna acceptera att butiker skapar kundprofiler baserade på videoövervakning i kombination med användning av Wi-Fi-nätverk, inloggningsinformation, kreditkortshistorik, och så vidare? Vilka kampanjer är acceptabla och vilka är på gränsen? Hur bör eventuella regler utformas?

Om vi bortser från att dagens praxis at läkare i dag inte gör uppskattningar av patienters livslängd av praktiska och etiska skäl: Om uppgifter i medicinska journaler av en patient indikerar att han eller hon har en förväntad livslängd på 12 månader - baserat på statistiska analyser av ett tillräckligt antal patienter med motsvarande registrering - ska patienten veta om det? Vad händer om journalen är tillgänglig för patienten? Hur kommer detta att påverka hur läkarna behandlar patienten? Bör hälsomyndigheterna kunna sätta stopp för särskilt dyra behandlingar?

Om en persons profil matchar beskrivningen av en brottsling i en sådan utsträckning att det är mycket troligt att polisen förutspår att personen i en nära framtid kommer att begå en kriminell handling, bör polisen ingripa? Hur? I så fall bör man skilja mellan olika typer av handlingar (bedrägeri, bankrån, terroristhandlingar, och så vidare)? Ska polisen låta ärendet vara eller är det acceptabelt att polisen använder förebyggande åtgärder som allvarliga samtal, eller borde de vidta allvarligare åtgärder? Och vad händer om personen inte har godkänt att polisen kan inneha informationen som ger den relevanta beslutsgrunden - ska den potentiellt kriminelles integritet och "oskyldig tills motsatsen har bevisats" väga tyngre än att skydda ett potentiellt offer?

Andra intressanta frågor:

- Bör polisen kunna stänga av vägar och områden där det statistiskt sett kan förväntas uppror, till exempel vid sportevenemang eller besök av statsöverhuvuden?
- Ska försäkringsbolagen kunna justera priser och villkor baserat på din profil?
- Bör banken kunna justera lånevillkoren baserat på din profil?
- Bör potentiella arbetsgivare utvärdera kandidater baserat på big data?
- Om hälsomyndigheterna ser att någon väntar barn men inte är lämplig att ge barnet en tillräckligt bra uppväxt, borde de ingripa?

De flesta av dessa frågor har inte ett tydligt, svartvitt svar. Det är inte svårt att argumentera och åsikter kommer att variera från person till person. Eftersom big data gör intåg kommer vi att utmanas både vad vi tycker verkar vara bra och hur lagen fungerar.

32. "How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did", Forbes, 16. februar 2012

Praktisk tillämpning:
Applikationer och exempel

Devices och “the Internet of Things”

Smartphones registrerar enormt mycket information som geografisk plats, datatrafik, applikationsanvändning, bilder, video, ljud, rörelse och mycket mer. Totalt finns det cirka 4,5 miljarder mobilabonnenter världen över, och en fortfarande växande andel av dessa använder smartphones: I början av 2014 uppskattade IDC att cirka 63% av alla världens mobiltelefoner är smarttelefoner (upp från 50% föregående år) och att den genomsnittliga smarttelefonen nu bara kostar $316 USD (33).

Med andra ord, fler och fler människor runt om i världen kommer att lämna ut mer data under de kommande åren. Dessutom kommer "the Internet of Things" bidra till ständigt ökande datamängder - när "wearables" (GPS-objekt, klockor, glasögon, kläder), bilar, kylskåp, TV-apparater, hus och vem-vet-vad, kommer att vara online och generera data. Det är också snack om hur "smarta hus" kommer att kunna ansluta allt i ett enda stort nätverk: Om det finns en brand i huset kommer brandlarmet att meddela brandkåren, brandbilen dirigeras det snabbaste vägen till huset medan annan trafik avleds, vattenförsörjningen till området säkras, försäkringsbolaget informeras och så vidare - allt automatiskt.

Datatrafik

Leverantörer av internet- och mobildata (Telenor, Netcom, Canal Digital, Get osv.) vidare) vet mycket om beteendet hos enskilda användare; vad de surfar på, hur mycket data de konsumerar, när, fysiskt var, och så vidare. Till exempel ser man att datatrafiken kraftigt belastar det lokala nätverket på Fornebu när stora konserter hålls på Telenor Arena, och man kan se hur abonnenter rör sig utifrån vilken basstationer de är associerade till vid olika tidpunkter. Dessutom vet leverantörerna förmodligen vilken typ av enhet varje abonnent använder och vilken typ av applikationer och data som konsumeras, liksom vem och hur de kommunicerar med andra (SMS, samtal). Denna information krävs att leverantörerna raderar efter tre månader (34) (sex månader om datalagringsdirektivet införs); om man kan lita på att tjänsteleverantörerna följer lagstiftningen är en annan fråga.

Sociala medier

Facebook, Twitter, LinkedIn, Instagram, Snapchat och en mängd andra sociala medier hjälper användare att publicera stora mängder personlig information varje dag. Vi hör regelbundet om hur NAV använder Facebook att utsätta drogmissbrukare (35). Facebook hamnade nyligen i negativt ljus då visade sig att de 2012 hade genomfört ett experiment med nästan 700 000 användare för att se om de kan manipulera användarens känslor (36). Dessutom ökade medvetenheten om att Facebook tydligen kräver ökad insyn och tillgång till enskilda användares enheter när du använder IM-tjänsten Messenger, även om Facebook själv förklarar att det inte är så illa som det verkar (37).

Google Trends

Det finns många söktjänster, men Google är utan tvekan ledande, med 90% världsmarknadsandel (38). Det finns många berättelser om hur Google kan förutsäga händelser baserat på förändringar i sökbeteenden; till exempel kan vissa nyckelord vara bra ledtrådar till influensaaktivitet (39). Även om grad av noggrannhet är diskutabel, det finns också de som tar upp att Google Trends kan används för att förutsäga hur finansmarknaderna kommer att röra sig (40). Twitter-aktivitet kan också - bland annat - användas för samma syfte som Google Trender.

33. «Smartphone OS Market Share, Q1 2014», IDC
34. "Så mye er lagret om deg", Mobilen, 4. juni 2012
35. "- Nav bruker avhørsmetoder for å avsløre juks", Eirik L. Berglund og Stella Bugge, E24, 25. mai 2012
36. "Furor Erupts Over Facebook's Experiment on Users", Reed Albergotti, Wall Street Journal, 30. juni 2014
37. "Facebook Messenger Privacy Fears? Here's What to Know", Albergotti, R. Wall Street Journal, 8. august 2014
38. "Yahoo slår Google", Computerworld, 23. august 2013
39. Google.org Influensatrend
40. "Big Data Gets Bigger: Now Google Trends Can Predict The Market", Forbes, 26. april 2014

Köpbeteende och CRM (customer relationship management)

Amazon, Netflix, Spotify är tre exempel på företag som använder betydande resurser för att förstå och använda kundbeteende. Amazon ger rekommendationer av typen "läsare som köpte X, köpte också A, B och C", medan Netflix och Spotify rekommenderar och marknadsför filmer och låtar baserat på vad du har sett och hört innan de kombineras med data om hur användare med preferenser matchar din har också sett och hört. De flesta onlinebutiker har också stora mängder kund och köpinformation som enkelt kan kombineras med andra typer av datauppsättningar.

Privatekonomi

Folkbokföringen, skattedeklarationer, skatteinformation, kreditbedömningar och bank-/ kreditkortstransaktioner innehåller alla värdefulla personliga finansieringsuppgifter till människor både ensamma och totalt. Lojalitetsprogram som gör det möjligt att få djupare inblick i konsumenternas shoppingmönster, och denna insikt lönar sig i form av 1% återbetalning av det totala inköpspriset.

Finans och mikroekonomi

I Norge (och många andra länder) krävs företag och andra organisationer att tillhandahålla ett minimum av redovisningsinformation (inkomst, kostnader, inkomst, skuld, eget kapital, utdelning etc.) på årsbasis. Kombinerad med annan information om organisationer som organisationsform, organisationsnummer, bransch, adress, antal anställda, ägarstruktur, ägare, ledning, aktiekapital, grundande datum, nedläggningar, konkurser och mer är lätt att lära sig mycket om näringslivet. Exempel: www.allabolag.se

Finans och makroekonomi

Kvartals- / årsrapporter, aktier, fonder, valutor, obligationer, derivat, råvaror, räntor, inflation, sysselsättning, inhemsk produkt, handels-/import- /exportera och en hel mängd annan information om hur internationella och nationella ekonomier rör sig är till stor del tillgängliga för allmänheten. Exempel: http://finance.google.com.

Annan offentlig statistik

Folkbokföringen, skolbetyg, brottsstatistik, hälsa, social trygghet, statistiska centralbyrån, trafik och tidtabeller (bilar, tåg, buss, spårvagn, flyg, båt/fartyg), invandring, utvandring och så vidare. Som exempel erbjuder New York City över 1 300 olika datauppsättningar gratis (41). Flera norska direktorat?? börjar också bli bättre på att släppa data.

Väderdata

Väderdata är kanske den äldsta av big data - enligt Norwegian Meteorological Institutet (MET) är Norges äldsta väderdata från 1761 (42). Världen över registreras väder- och havsdata fortlöpande och genom ett antal initiativ är målet bland andra MET som meteorologiska, oceanografiska, hydrologiska och klimatdata bör kunna vara kompatibla med andra filformat för utbyte mellan system. Varje dag måste MET ta emot, bearbeta och göra hundratals GB data tillgängliga i olika format på extremt kort tid. Väderappen som de flesta av oss har på smarttelefonen återger information baserad på enorma och snabba förändringar mängder data.

41. NYC Open Data Plan Dashboard
42. “Databaseutvikling”, Norsk Meteorologisk Institutt

Navigering och trafikledning

Med Google Maps kan du till exempel se hur trafikflödet går igenom de viktigaste genomfarterna i de flesta städer. Baserat på förväntad trafikbelastning (baserad på historiska data) eller verklig trafikbelastning (baserad på realtidsdata) kan man få förslag till alternativa rutter. När du planerar träningsrutter i Strava kan du välja populära sträckor baserat på var de flesta GPS-spår finns. När man i framtiden får självstyrande bilar, som Google för närvarande undersöker (43) (i april 2014 har de totalt kört över 1 miljon kilometer), kommer man att kunna kombinera "the internet of things" med big data för att få effektivare trafik utan involvering av människor. Om ett utryckningsfordon måste åka till en destination, kan trafiken styras så att fordonet anländer så snabbt som möjligt, till exempel genom att varna trafikanter, dirigera om bilar till sidan osv.

Strava Metro (44) är en tjänst från Strava, som är mer känd som en GPS-baserad träningsdagbok för cyklister och löpare. Varje vecka laddas det upp 2,5 miljoner sessioner GPS-spår runt om i världen (antalet i Norge vill de inte att ange), och enligt Strava är cirka hälften av alla registrerade sessioner relaterade till pendling. Världen över har Strava-användare registrerat cirka 80 miljoner cykelturer och 21 miljoner löpturer, totalt över 220 miljarder datapunkter. Strava har därför en unik insikt i hur cyklister pendlar in och ut ur städer, insikten är också tillgänglig via Strava Metro service. Detta är information de säljer, och de som planerar infrastruktur i städerna kan dra stor nytta av att få detaljerad information om varför pendlare rör sig som de gör (45).

Tillgängligt för alla

Det är inte bara de stora organisationerna och företagen som kan dra nytta av big data: eftersom data och datakraft har blivit billigare, mer användarvänligt och mer tillgängligt kan också små och medelstora företag använda affärsinformation och big data. Från New York Times (46) kunde vi läsa om hur en biltvätt använde teknik och offentlig tillgänglig big data för en bättre förståelse och effektivare drift av verksamheten.

Kombination av data

En sak är när man använder stora datauppsättningar för att lära sig något om ett visst ämne, men big data blir riktigt intressant när du kombinerar två eller flera datamängder. Ingen datauppsättning är den andra lik, så arbete i varierande grad krävs för att kunna slå samman dem.

Sammanslagning av datauppsättningar för att skapa insikter kan starta top-down eller bottom-up:

- **Top-down:** Problem som ska besvaras
- **Bottom-up:** Förklarande; upptäcka relationer med hjälp av experimenterande

43. Wikipedia-artikkel om Googles selvstyrte biler
44. Strava Metro
45. http://blog.visma.com/no/big-data-personvern-og-innsikt/
46. «Finding Affordable Ways to Use Big Data», New York Times, Juli 2014

Här är några exempel - som antagligen kan byggas ut i det oändliga:

Dataset #1	Dataset #2	Sammanhang	Åtgärder
Google Maps	MET	Hur hänger vädret samman med trafikförhållande/framkommlighet?	Aveckling ac trafik, förhindra olyckor
Netflix	MET	Hur hänger vädret samman med konsumering av filmer/serier?	Annonseringstillfällen för Netflix
Google-sök (Google Trends)	Finansiell data	Vilka sökord/-fraser upprepas mest inom finansmarknaden?	Investeringstips
Google-sök (Google Trends)	Sjukdomsstatistik	Vilka sökord/-fraser upprepas mest i influense- och epidemiutbrott?	Förbereda akutmottagningar, vaccinering
Spotify	Netflix	Vilka filmer gillar de som lyssnar på en specifik typ av musik?	Öka användandet av musik och film
Sportresultat	Kriminalstatistik	Hur hänger våldsbrott samman med resultaten för den lokala fotbollsklubben?	Öka polisbemanning i kritiska områden
Nybilsförsäljning	Omsättningssiffror för restaurangbranschen	När människor har råd med nya biler har de oftast råd till att äta mer ute	Öka resurser/bemanning i restauranger
Skatt	Kriminalstatistik	Hur är sammanhanget mellan föräldrarnas intäkt och barns prestation i skolan?	Prioritering av lärare till regioner med större problem
Personstatistik	Kriminalstatistik	Klargör eventuellt sammanhang mellan prestationer i skolan och kriminalitet	Prioritering av lärare till regioner med större problem
Skatt	Kreditupplysningnar	Klargör sammanhanget mellan betalningsförmåga och ansöknader om lån	
Facebook	Skattelistor	Se hur dina vänner ooch bekanta ligger till lönemässigt	Snoka

*) En bekant till mig som driver flera restauranger på Aker Brygge berättade att detta är en vanlig indikator i deras bransch.
Antagligen kan nybilsförsäljning ersättas med en annan underliggande drivkraft som BNP, löneökning eller andra KPI:er

Teknik, lösningar och kompetens

Hittills har detta dokument tagit upp vad big data är på en övergripande nivå och vad det kan användas till. Allt detta är emellertid inte möjligt utan några av följande underliggande faktorer på plats: big data kräver teknik i form av lagringsmedier, databaser och programvara på olika nivåer som tillsammans representerar en lösning. På grund av den potentiellt komplicerade komplexiteten, behövs även kompetenta människor till att möjliggöra big data - vissa organisationer har det som behövs internt, men det kommer antagligen vara fördelaktigt att ta in extern expertis. Detta utgör tillsammans ett ekosystem av leverantörer av hårdvara, mjukvara och tjänster.

Lagringsmedia

Tidigare lagrades data på disketter (1,4 MB per disk) - nu är till och med små företag uppe i flera GB data, och de största företagen i världen är uppe i petabytes och exabyte av data (1 EB = 1 000 PB, eller 1 000 000 000 GB). Netflix filmdatabas består av 3,14 PB, vilket motsvarar 17 miljoner kilo disketter. Om du skulle överföra ditt Netflix-bibliotek till 3,5-disketter, skulle du behöva lägga nästan en miljard disketter bredvid varandra, ett avstånd på 2 083 km - motsvarande Oslo till Rom.

Mängden (affärs)data i världen antas fördubblas var 15: e månad och vi genererar 2,5 exabyte per dag idag (nästan 700 miljarder disketter), vilket är mer data än någonting som produceras över hela världen från tidiga tider till 2003.

Se Gartners 3V-definition: big data kräver snabb hantering av stora datamängder; därför är lagringsmedium kritiskt. Med ständigt ökande volym behövs ny teknik och nya metoder för att uppfylla nya krav på tillgänglighet och respons.

Å ena sidan har du den traditionella disken för att lagra databaser där du sparar och ändrar data genom att göra ändringar på en roterande disk med elektroniska, optiska eller mekaniska mekanismer (hårddisk, diskett, CD, etc.). Solid-state-enheter (SSD) omfattas också av denna definition. Två kritiska variabler för traditionella diskar är hur lång tid det tar att få tillgång till data (access time) och hur lång tid det tar att överföra data (transfer time).

Å andra sidan har man in-memory databaser, som har blivit mer vanligt allt eftersom priset på RAM har sjunkit. När databasen finns i minnet kommer frågor att utföras snabbare, delvis på grund av att algoritmerna är enklare och CPU:erna behöver färre instruktioner. Denna lösning är utbredd när responstiden är kritisk, till exempel inom telekom- och mobilreklam. En hybridlösning till detta är att cacha information. Ett bra exempel är HydCache, en lösning som utvecklas vid norska vattenresurs- och energidirektoratet (NVE). HydCache hämtar data från olika platser (NVE:s interna tjänster såväl som vissa externa tjänster som Metrological Institute), ritar diagram baserat på platserna. Till exempel kräver en graf 10 frågor (internt och externt), vilket tar tid att köra - särskilt eftersom responstiden är långsam. Uppgifterna är relativt statiska (tidsserier med uppmätta värden). Lösningen hämtar data från alla platser och lagrar dem i en databas i minnet och graftjänsten hämtar data från denna databas.NVE använder in-memory-kapaciteten i Microsoft SQL 2014-lösningen. En annan möjlig lösning är en distribuerad minnescache ovanpå SQL-servern. Mängden data ligger i storleksordningen 40-50 miljoner rader, beroende på hur mycket som ska cachas.

Det finns lösningar som är en hybrid av konventionell disk och in-memory, till exempel PCi express. Detta ligger dock utanför omfattningen av detta dokument.

Databaser

Traditionellt har data lagrats i relationsdatabaser, där data är strukturerade och lagras i ordnade former. Detta fungerar bra med mindre datauppsättningar men är inte särskilt skalbar: när datamängden ökar till den volym som man har idag, och datan är ostrukturerad, möter man begränsningar med vanliga databaser. När mängden ökar sjunker prestandan. De klassiska exemplen är Microsoft SQL och MySQL. Som en lösning på den begränsning som följer med relationsdatabaser har NoSQL setts som en lösning. NoSQL står för "Not only SQL ». Istället för att data struktureras på ett visst sätt, är databasers hårda krav nedgraderade för att få prestanda; det vill säga att NoSQL kan hantera både strukturerad och ostrukturerad data. Databaserna är ofta spridda över många noder och servrar och man kan lägga på mer hårdvara kontinuerligt. NoSQL-databaser kan kategoriseras enligt följande:

- Key-value stores använder så kallade associativa array som datamodell, där nycklar lagras i par (nyckel + värde/uppslag). Vissa lagrar data i minnet (RAM) medan andra lagrar data på vanliga hårddiskar. Exempel inkluderar Amazon DynamoDB, Memcached, Coherende och Redis.

- Dokument-orienterade stores, som organiserar dokument efter samlingar/ collections, tags, metadata och kataloger (directories). Några exempel på databaser: Apache CouchDB, ElasticSearch, IBM Notes, IBM Domino, MarkLogic Server, MongoDB (mest populära), Oracle NoSQL-databas, SimpleDB, Cloudant med flera.

- Graf-orienterade databaser, där data kopplas till varandra genom relationer vilket bäst beskrivs grafiskt - till exempel sociala relationer (vänner, kollegor), tåglinjer (stationer), kartor osv. Några exempel på databaser: AllegroGraph, FlockDB, IBM DB2, InfiniteGraph, Neo4j, OrientDB, Zones GraphDB, Sqrrl Enterprise, Stardog med flera.

- Kolumn-baserade databaser lagrar datatabeller i kolumner istället för rader (relationella databaser använder vanligtvis rader). Detta är en fördel för till exempel datalager och CRM-system, där stora mängder data ofta aggregeras. Exempel inkluderar Google BigTable, Hbase och Accumulo.

- RDF-butiker, även känd som triplestore / resursbeskrivningsram (RDF = resource description framework): Entity-attributebalue-modell (Enhetsattribueringsmodell), där syftet är att skapa en mekanism för kodning och tolka resurser på ett sätt som en specifik programvara kan förstå.Användbart för att identifiera vem som känner vem i ett socialt nätverk. Används också för att lära sig mer om trafikmönster. Exempel inkluderar Apache Jena och Sesame.

- XML-databaser: Lagrar data i XML-format som du kan köra frågor på, exportera och omformatera till önskat format. Används vanligtvis med dokumentorienterade databaser. Exempel inkluderar BaseX, Sedna och eXist.

- I objektdatabaser är information tillgängliga som objekt.? Objekt är särskilt väl lämpade för multimediainformation. Exempel inkluderar XODB, DB40, Eloquera, Versant, Objektivitet DB och Velocity DB.

- Hierarkiska databaser organiserar data i en trädstruktur. Data lagras som poster och är länkade på samma sätt med länkar; en "parent" kan ha n barn, medans ett barn kan bara ha en "parent". Detta avslöjar också begränsningen i "en till många"-förhållandet som databasen skapar. Lämplig för att hålla ordning på anställda i en standardstruktur.
  Exempel inkluderar InterSystems Cache och GT.M.

- Multimodell databaser. Exempel inkluderar arangodb, Datomic, Orient DB, FatDB och AlchemyDB.

Några välkända exempel på NoSQL, både för on-premis installation och molnbaserad:

- Amazon SimpleDB
- Cassandra
- Google AppEngine
- MongoDB
- MongoLab
- Redis
- SalesForce Database.com

Software

För att hantera informationen i de olika databaserna behövs programvara av olika slag.

Molnbaserade tjänster kan drivas av Amazon Web Services, Amazon EC2, RackSpace, Windows Azure, Google Compute Engine, Joyent, Heroku, Cloud Foundry, etc. Fördelarna med molnbaserade tjänster är många och inkluderar förvaltning, drift, uppdateringar och säkerhet tas helt eller delvis hand om av tjänsteleverantören. Nackdelarna inkluderar mindre kontroll och risken för obehörig åtkomst.

On-premis tjänster måste installeras och hanteras av individen, men kan samtidigt tillhandahålla bättre säkerhet.

Vad du väljer beror naturligtvis på ett antal faktorer som vikten av data, budget, intern kompetens etc.

Hadoop (aka Apachee Hadoop) är inte en databas utan ett ekosystem av mjukvara designad för parallell-behandling av data. Bland annat Hadoop NoSQL. Hadoop använder också MapReduce, som är en modell för att sprida databehandling över olika servrar (= Hadoop-kluster) så att processorkraften hos flera servrar används samtidigt - vilket minskar kraftigt vid bearbetningstiden. Hadoop kan installeras på lokala servrar eller på servrar i molnet. Av molntjänstleverantörer som erbjuder Hadoop är den mest kända Microsoft, Amazon (Elastic Compute Cloud (EC2), Simple Storage Service (S3) och Google.

Hadoop innehåler följande moduler:

- Hadoop Common: Bibliotek och verktyg för andra Hadoop-moduler.
- Hadoop Distributed File System (HDFS): Filsystem som lagrar data på maskiner, med hög bandbredd över klustret. Filer eller datauppsättningar finns vanligtvis i storleksordningen gigabyte eller terabyte.
- Hadoop YARN: Resurshanteringsplattform för hantering och planering användning av resurser
- MapRecude (Hadhoop MapReduce): Programmeringsspråk/ramverk för att processa parallellt genom att använda flera datorer (noder), även kallat cluster/grid.

Kända användare av Hadoop inkluderar Yahoo, Facebook (beräknas vara 500 PB per i idag, ökar med 0,5 PB dagligen) och en hel mängd företag på Fortune 500-listan.

Kompetens

Utan duktiga människor kommer man dock inte lyckas med big data. "Data scientists" (47) - är ofta nyfikna människor som både bemästrar programmering och statistik - de är viktiga resurser för framgången för investering inom big data. Kort sagt det krävs en god förståelse för data och IT, statistik och affärsverksamhet (48). HR-avdelningen kommer förmodligen att bekräfta att personer som har dessa tre egenskaper är unika och förmodligen inte lätta att rekrytera. Det är värt att notera att det finns en väsentlig skillnad mellan "data scientists" och analytiker (49); de förstnämnda är bättre på att se hur deras bidrag är relaterad till övergripande verksamhet och kan lättare engageras av komplicerade frågor, men är inte lika bra på att kommunicera sina resultat med sin omgivning. För att lyckas med att dra nytta av deras kompetens krävs engagemang från företagsledningen. Mer specifikt pekar litteraturen på följande åtgärder: Höj analysen till ledningsnivå, se till att beslutsfattare stöttar upp under arbetet och arbetar nära "data scientists". Ge "data scientists" tillräckligt med frihet att ta tag i problemen samtidigt som de utmanas att i första hand få kunskap om de data de arbetar med (t.ex. prata med slutkunder), sätta ihop ett team av "data scientists" med varierande bakgrund och expertis och ge dem incitament som betyder något för dem.

47. Davenport, T.H. & Patil, D.J. (2012), "Data Scientist: The Sexiest Job Of the 21st Century", HBR, October 2012
48. Davenport, T.H. (2013), "Keep Up with Your Quants", HBR, July-August 2013
49. Harris, J.G. & Mehrotra, V. (2014), "Getting Value From Your Data Scientests", MIT Sloan Review, Fall 2014