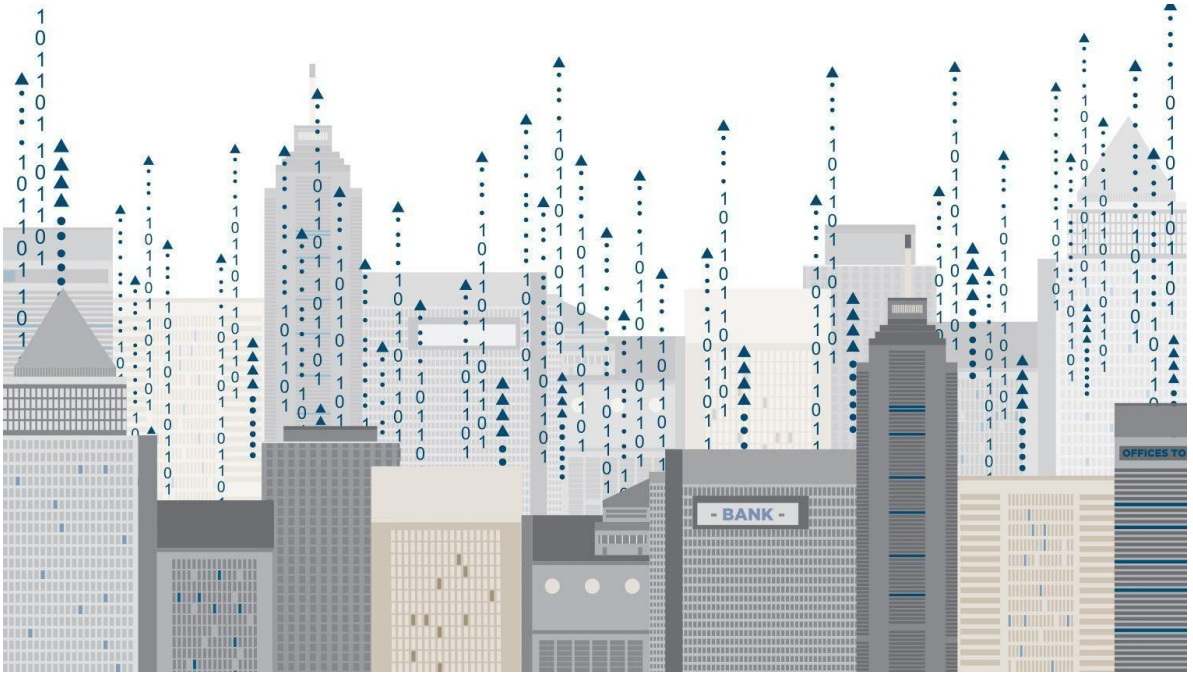
BEST PRACTICES BIJ BIG DATA ANALYTICS

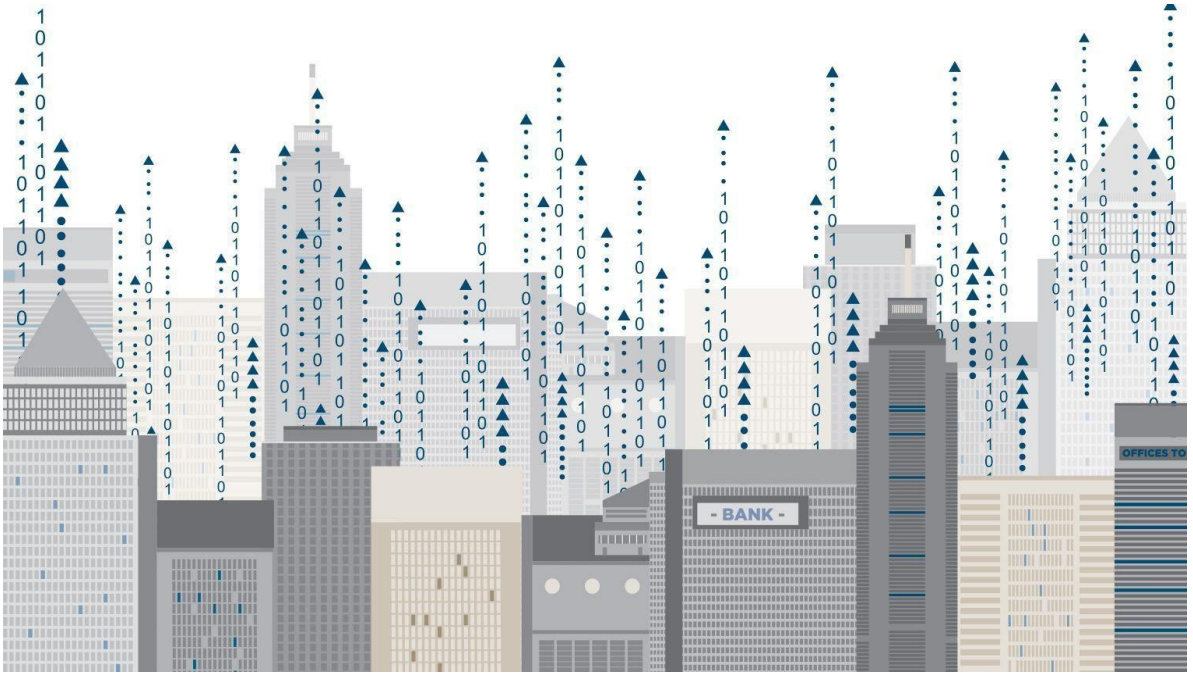
Door Jirry Haerinck, Wouter Duckaert, Ken Bostoen en Jorick Triempont



2016-2017
Karel de Grote-Hogeschool Antwerpen, Handelswetenschappen en Bedrijfskunde
Groep 09

BEST PRACTICES BIJ BIG DATA ANALYTICS

Door Jirry Haerinck, Wouter Duckaert, Ken Bostoen en Jorick Triempont



2016-2017
KAREL DE GROTE-HOGESCHOOL ANTWERPEN, HANDELSWETENSCHAPPEN EN BEDRIJFSKUNDE
Groep 09

Voorwoord

Deze paper werd geschreven door vier laatstejaarsstudenten toegepaste informatica van de Karel de Grote-Hogeschool in Antwerpen. De inspiratie voor het gekozen onderwerp was de 'big data analytics'-presentatie van Tine Van Dyck op de TI-conference.

Wij danken de volgende personen en bedrijven: onze coach Hans Vochten, die steeds klaarstond met advies en antwoorden op onze prangende vragen; docent Jan Van Overveldt, die ons in contact bracht met heel wat bedrijven, waaronder Big Industries, Infofarm, I4BI, SAS en Deloitte; de personen op deze bedrijven die ons in staat stelden tot het afnemen van interviews (Deloitte: Els Renders; SAS: Elisabeth Versailles, Tine Van Dyck, Anthony Severeyns; Big Industries: Matthias Vallaey; Infofarm: Ben Vermeersch; I4BI: Hannes Cassiers, Maarten Van den Broeck). Extra dank gaat uit naar SAS, voor het ter beschikking stellen van een licentie voor het opzetten van een demo-project.

Inhoudsopgave

1. Inleiding

1.1 Situering

Het belang van big data voor bedrijven allerhande valt dezer dagen niet te overschatten. De capaciteit om grote hoeveelheden ongestructureerde data om te vormen tot inzicht, wordt stilaan een noodzakelijk gegeven dat voor een groot stuk bepaalt hoeveel klanten een bedrijf kan vergaren, hoe goed het deze klanten kan bedienen en bijgevolg of het bedrijf in kwestie überhaupt overeind kan blijven. Weten hoe je een systeem moet opzetten om big data te analyseren, wordt dus cruciale kennis voor een snel toenemend aantal bedrijven.

1.2 Probleemstelling

Het opzetten van een big data analytics systeem is echter geen evidente opgave. Hierbij moeten verraderlijke valkuilen worden vermeden en advies van ervaren experts moet worden opgevolgd. Aangezien er momenteel echter een nijpend tekort is aan experts in het veld van big data analytics, en dit tekort de komende jaren wellicht alleen maar zal toenemen, kan niet elk bedrijf rekenen op het advies van een ervaren expert. Daarenboven moeten deze experts zelf ook hun sporen verdienen en leren door scha en schande - een proces dat veel tijd vergt.

Een bevattelijk corpus van best practices bij big data analytics, geconstrueerd op basis van het advies van ervaren experts, zou een mogelijk hulpmiddel kunnen zijn voor onervaren experts om sneller up-to-speed te geraken, en voor niet-experts om toch een verdienstelijke poging te wagen om op beperkte tijd voldoende te leren om een degelijk big data analytics platform op te zetten.

Met deze paper willen we een poging wagen om een dergelijk corpus van best practices te construeren. Uiteraard zijn wij zelf echter geen ervaren experts in dit domein. Daarom zullen we te rade gaan bij ervaren experts, en dit via de twee volgende wegen: enerzijds vanuit literatuurstudie en anderzijds vanuit het afnemen van interviews.

1.3 Opbouw van de paper

In een eerste beweging van het schrijven van deze paper, zullen we een aantal best practices oplijsten op basis van de research die we verrichten. Een eerste soort bron hiervoor zal zijn wat we terugvinden in literatuur, d.w.z. in white papers, en in artikels uit kwaliteitsvolle magazines of op toonaangevende websites. Parallel met deze literatuurstudie, zullen we onze boter gaan halen bij een tweede soort bron, namelijk interviews met lokale experts. Welke de best practices zijn die we uit deze twee soorten bronnen distilleren en toevoegen

aan ons corpus, zal deels afhangen van het aantal bronnen waarin we deze best practice kunnen terugvinden, deels van hoe goed en hoe uitgebreid het bestaan van deze practice wordt beargumenteerd en toegelicht in de bron in kwestie.

De logische volgende stap is dan het samenvoegen van de best practices die we in de twee voorgaande stappen hebben geformuleerd. Dit samenvoegen zal in eerste instantie een vrij intuïtieve onderneming zijn (tenzij we instanties tegenkomen waarin de best practices uit het eerste luik ingaan tegen de best practices die we hebben geformuleerd in het tweede luik). In tweede instantie, echter, zal dit samenvoegen nog iets verder gaan dan het louter aan elkaar kleven van deze twee lijsten. Het is immers de bedoeling om de best practices die we hebben gevonden op een bevattelijke, meer systematische manier voor te stellen.

Onze finale stap in de theoretische component van deze paper zal bijgevolg zijn om meer structuur te brengen in onze lijst met best practices, deels door verschillende best practices te groeperen onder eenzelfde noemer, deels door visuele modellen te construeren die de samenhang van deze best practices verduidelijken.

Het laatste deel van onze paper zal dan bestaan uit een praktische component, waarin we de best practices die we hebben geformuleerd, toepassen in een project.

2. Begrippen verder uitgelegd

2.1 Big data

2.1.1 Wat is big data [1]

*Figuur SEQ Figuur * ARABIC 1 Big Data*

Big data kunnen we voorstellen aan de hand van een stroom van verschillende gegevens. Voordat we bepaalde gegevens onder de noemer van 'big data' kunnen zetten, moeten ze voldoen aan enkele definities. Deze definities worden ook wel de 3 V's genoemd (Volume, Velocity en Variety).

Volume

Volume komt het meest aan bod als men over big data spreekt. Dit komt omdat data steeds groter wordt en gigantische volumes kan krijgen. We zien de laatste tijd een exponentiële groei in data storage. Een van de grote redenen hiervoor is dat de data die opgeslagen wordt vaak andere dingen zijn dan puur tekst. Er worden tegenwoordig bijvoorbeeld veel video's, afbeeldingen of muziek opgeslagen op social media.
Om al deze data te beheren en zinvol te maken, is het concept van 'big data' ontstaan.

Velocity

Vroeger werd data in stukken geanalyseerd en verwerkt. Bedrijven namen een stuk van hun opgeslagen data, voerden daar analyses op uit en keken naar het resultaat. Deze manier van werken gaat enkel als de data op een traag tempo gegenereerd wordt. Door alle sociale media en de snelheid van data die aangemaakt wordt, moeten de analyses hier ook op aangepast worden. Men moet alle data real time verwerken; dit zorgt ervoor dat er veel sneller beslissingen genomen kunnen worden. Door de grote hoeveelheid aan data real time te verwerken, weet een bedrijf ook meteen welke beslissingen ze moeten nemen. Je kan dit vergelijken met filegegevens. Je wil niet weten dat het een uur geleden file was, maar op dit moment.

Variety

Als je data in grote hoeveelheden binnenhaalt is het ook belangrijk om te zien in welk formaat deze binnenkomt. Data die binnenkomt kan bijvoorbeeld heel gestructureerd zijn, waaraan je direct veel kan zien. Maar ze kan ook binnenkomen als ongestructureerde data. Een combinatie van gestructureerde en ongestructureerde data kan ook voorkomen.

---

[1] J. VAN OVERVELDT, *Big Data Introductie,* Antwerpen, Karel De Grote-Hogeschool, l.d., 6; SAS INSTITUTE NV/SA, http://www.sas.com/en_us/insights/big-data/what-is-big-data.html, geraadpleegd op 23 januari 2017.

Conclusie 3 V's

Big data is dus grote hoeveelheden van (on)gestructureerde data die aan hoge snelheid binnenkomt. Door constante vooruitgang/vernieuwing van applicaties en technologie ontstaan er nieuwe vormen van data.
Het concept van het verzamelen en opslaan van grote hoeveelheden van data om later te verwerken bestaat al zeer lang. Het was pas rond het jaar 2000 dat de term 'big data' een duidelijke definitie kreeg, Doug Laney heeft dan de drie V's beschreven. Deze V's zijn de basis van wat 'big data' juist betekent, maar veel mensen voegden hier later nog extra V's aan toe. Hierdoor spreken websites vaak over de 4 of 5 V's van big data.
Big data kunnen opslaan op zich heeft niet veel nut, het is belangrijk om te weten wat je er allemaal mee kan doen. Het juist interpreteren van al die data kan een gigantische impact hebben op het bedrijf. Hierbij komt dan het analytische aspect aan bod.

2.1.2 Gedistribueerde opslag en gedistribueerde verwerking [2]

Het is heel belangrijk om te weten hoe deze big data technologieën omgaan met de 3 V's. Dit doen ze door gebruik te maken van gedistribueerde opslag en gedistribueerde verwerking. Het concept van gedistribueerde verwerking is eigenlijk het verdelen van een groot probleem in verschillende deeltaken. Deze deeltaken worden dan weer op hun beurt verdeeld over verschillende nodes, workers of servers. Deze werking van opsplitsing is niet zichtbaar voor de eindgebruiker maar gebeurt wel. Hetzelfde geldt voor gedistribueerde opslag. Bij gedistribueerde opslag gaat er geen groot probleem opgesplitst worden maar grote dataobjecten zoals files of tabellen. Die worden op hun beurt ook verdeeld in deelobjecten en over verschillende nodes verspreid.

2.1.3 Dataopslag [3]

De data die je gaat binnenhalen met big data technologieën ga je natuurlijk ook ergens moeten opslaan. De standaard relationele databanken zoals Oracle en MySQL kunnen hier niet echt goed mee om. Doordat er heel veel data meestal ad-hoc verwerkt moet worden, treden er heel veel lees- en schrijfoperaties tegelijk op. Dit is een bottleneck waar de standaard relationele databases niet mee overweg kunnen. Om dit op te vangen ga je van SQL-databanken moeten overstappen naar NoSQL databanken of NewSQL databanken. Vroeger waren alle data sources relationele databases, werd door middel van data integration automatisch verschillende data uit de bronnen gehaald en in het datawarehouse gestoken. Het data warehouse kon toen bezien worden als centrale repository waar alle data geïntegreerd in zat. Hierop werden dan nog lagen opgezet om analytisch mee te gaan werken.
Door de komst van big data is dit systeem helemaal niet meer goed om mee te

---

[2] J. VAN OVERVELDT, *Big Data Introductie,* Antwerpen, Karel De Grote-Hogeschool, l.d.,7.
[3] COMPUTERWEEKLY, http://www.computerweekly.com/podcast/Big-data-storage-Defining-big-data-and-the-type-of-storage-it-needs, geraadpleegd op 23 januari 2017

werken. De data source is nu veel uitgebreider. De data is nu zo groot dat een databank veel te duur is, hierdoor wordt de data in een landing zone geparkeerd. Dit is eigenlijk een data lake[4]. Er worden wel tools ter beschikking gesteld zodat alle data in data lake beschikbaar is om te zien of er toegevoegde waarde uit gehaald kan worden. Om deze testen nu echt te gaan uitvoeren kan deze data naar een data warehouse verplaatst worden.

2.1.4 Hadoop ecosysteem [5]

De belangrijkste big data technologie die je tegenwoordig kan gebruiken is Hadoop. Dit is een platform waarop je alle zaken die je nodig hebt kunt gebruiken. Het meest bekende platform is Cloudera. Dit is een van de meest gebruikte platformen. Hierop kan je gebruik maken van Hadoop hdfs, een file systeem voor gedistribueerde opslag van gegevens. Maar je kan ook gebruik maken van Hbase en Giraph, beide NoSQL Databases. Om data processing uit te voeren kan je gebruik maken van Hadoop MapReduce, dit gaat de verwerking verdelen over verschillende nodes en tussenresultaten wegschrijven naar de disk. Een andere en betere manier is Spark, dit gaat dataverwerking grotendeels in memory laten doorvoeren en heeft een Streaming onderdeel dat gebruikt kan worden om data gestreamd te laten binnenkomen en realtime te verwerken. De belangrijkste databases zijn Spark SQL, Hive en Impala. Impala (en Tez) gaan hun code niet omzetten naar MapReduce taken en zijn daarom sneller dan Hive, die dit wel doet. Spark SQL zet zijn query's natuurlijk om naar Spark jobs. Als alternatief voor Cloudera hebben we ook nog Hortonworks. Wanneer een bedrijf echt volledig wil meegaan kan het ook in de cloud gaan werken. De belangrijkste cloud provider is natuurlijk AWS (Amazon Web Services).

---

[4] JAMESDIXON'S BLOG, https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/ , geraadpleegd op 23 januari 2017;
J. VAN OVERVELDT, *5_DataBeschikbaarMaken,* Antwerpen, Karel De Grote-Hogeschool, l.d.,8.

[5] APACHE HADOOP, http://hadoop.apache.org/, geraadpleegd op 23 januari 2017.

2.2 Analytics

2.2.1 Wat is analytics [6]

Het correct en snel analyseren van big data is zeer belangrijk voor bedrijven tegenwoordig. Om deze data te analyseren wordt meestal een bepaalde tool gebruikt. Enkele voorbeelden van tools zijn: R, SAS Visual Analyser, Tableau, … Door het gebruik van deze tools kunnen bedrijven big data omvormen naar interpreteerbare gegevens. Met deze gegevens kunnen ze dan verschillende beslissingen maken en voordelen behalen.

Het doel van big data analytics is om de interpretatie van grote volumes aan data gemakkelijk te maken. Door dit te doen kunnen nieuwe opportuniteiten en strategieën gevormd worden die gebaseerd zijn op inzichten. Deze kunnen in een bedrijf zorgen voor een toegevoegde waarde. Als een bedrijf deze inzichten juist gebruikt kunnen ze zo zorgen voor een competitief voordeel op de markt, en langdurige stabiliteit creëren.

Om het juiste resultaat te bekomen moet de juiste vorm van analyse gekozen worden. Er zijn meerdere vormen van analyse: descriptive, diagnostic, predictive en prescriptive.[7]

Figuur 3 Descriptive tot prescriptive

---

[6] J. VAN OVERVELDT, *1_Introductie,* Antwerpen, Antwerpen, Karel De Grote-Hogeschool, l.d., 4;
SAS INSTITUTE NV/SA, http://www.sas.com/en_us/insights/analytics/big-data-analytics.html, geraadpleegd op 23 januari 2017.

[7] J. VAN OVERVELDT, *1_Introductie,* Antwerpen, Karel De Grote-Hogeschool, l.d.,20 - 25.

Descriptive analytics

Deze analysetechniek draait rond het verklaren van wat er gebeurd is. Ze gaat aan de hand van gegevens een resultaat genereren zodat het bedrijf weet welke specifieke trends voorkwamen. Het zorgt voor inzicht in "het verleden".

Diagnostic analytics

Het draait hierbij rond "waarom is een trend voorgekomen". Het gaat een verklaring geven aan de hand van data mining en correlaties. Door deze analyse krijgen bedrijven een beter zicht op de data zodat ze de oorzaken van verschillende gebeurtenissen kunnen begrijpen.

Predictive analytics

Bij predictive analytics wordt de vraag "wat gaat er gebeuren" beantwoord. Door modellen en voorspellingstechnieken wordt er een beeld geschept van toekomstige trends. Aan de hand van deze resultaten kunnen bedrijven een idee krijgen van wat de evolutie naar de toekomst voorstelt.

Prescriptive analytics

Door het gebruik van prescriptive analytics kunnen bedrijven een antwoord krijgen op de vraag "Wat zouden we moeten doen". Het maakt gebruik van de predictive models om zo een idee te geven voor wat de volgende beslissing kan zijn.

2.2.2 BI-project [8]

Om een business intelligence (of big data analytics) project op te starten zijn er 3 fases.

Als eerste heb je de analyse van de business noden. Je moet in deze fase gaan bepalen welke beslissingen een business users nemen, wat hun doelen en objectieven zijn en in welke informatie ze meer inzicht willen krijgen.

Als tweede kom je bij de belangrijkste fase namelijk informatie ter beschikking stellen voor analyse. In deze fase ga je een selectie van je bronnen nemen, de datakwaliteit hiervan onderzoeken, indien nodig verschillende bronnen samen integreren, een data governance project opzetten en opleveren in het juiste en gewenste formaat. Dit laatste is belangrijk omdat operationele gegevens veel waardevolle analytische informatie bevatten maar er veel problemen optreden als je geen kennis hebt van de gegevens.

[8] J. VAN OVERVELDT, *1_Introductie,* Antwerpen, Karel De Grote-Hogeschool, l.d., 7 – 17.

Als laatste fase komen we terecht bij analyse met BI-applicaties. In deze fase ga je op basis van de beschikbare tools analyses uitvoeren om een toegevoegde waarde voor het bedrijf te bepalen.

2.2.3 Analytic Cycle for BI [9]

Om analyses uit te voeren ga je gebruik maken van een typische cyclische beweging, de 'analytic cycle for BI'. Deze cyclus bestaat uit activiteiten opvolgen (de stap waar management achter vraagt, maar ze hebben eigenlijk de andere stappen ook nodig), uitzonderingen identificeren, causale verbanden vaststellen, alternatieven modelleren en actie ondernemen en resultaat opvolgen. Om een goede BI-omgeving te hebben moet je de cycle ondersteunen en gebruikers aanmoedigen om deze te gebruiken.

2.2.4 Data Governance [10]

Een belangrijk aspect van analytics is data governance. Dit is een systeem waarin je gaat bepalen wie er verantwoordelijk is voor welke data en hierover kan beslissen. Dit gebeurt door gebruik te maken van vastgelegde modellen die bestaan uit een verzameling van regels. Maar ze beschrijven ook wie welke acties kan ondernemen met welke data, wanneer en voornamelijk onder welke omstandigheden. Dit met het doel om de consistentie te verhogen en vertrouwen in beleidsbeslissingen te ondersteunen. Het doel is natuurlijk ook om de inkomsten uit beschikbare data te verhogen en de efficiënte van het werken te verbeteren. Belangrijke begrippen bij data governance zijn verantwoordelijkheid, transparantie, auditeerbaarheid en 'stewardship'.
Met auditeerbaarheid bedoelen we dat alle beslissingen en processen aanpassingen moeten bijhouden van wie deze heeft doorgevoerd.
Met 'stewardship' wordt bedoeld dat er data stewards moeten zijn die de taak hebben om te controleren of data gerelateerde taken worden uitgevoerd met respect tot de governance regels.

---

[9] J. VAN OVERVELDT, *1_Introductie,* Antwerpen, Karel De Grote-Hogeschool, l.d.,27 – 36.
[10] J. VAN OVERVELDT, *3_DataGovernance*, Antwerpen, Karel De Grote-Hogeschool, l.d.,2-15;
DATA GOVERNANCE INSTITUTE, http://www.datagovernance.com/the-basic-information/, geraadpleegd op 23 januari 2017.

2.2.5 Master Data Management [11]

Met 'master data' bedoelen we data die referentie biedt naar sleutel entiteiten waarnaar vanuit processen verwezen kan worden. Dit kan bijvoorbeeld een gebruiker zijn. Met 'master data management' bedoelen we de methodiek die belangrijke referentie data op een centrale plaats ter beschikking gaat stellen voor heel de onderneming. Zo gaan we bronnen kunnen linken met elkaar en achterhalen welke overeenkomende waarden er in de verschillende bronnen zitten. Het bouwen van een master data management systeem is een data governance taak die ervoor zorgt dat er een betere datakwaliteit gaat ontstaan. Dit heeft ook als voordeel dat je bij het opkuisen van data minder werk gaat hebben.

2.2.6 Informatie ter beschikking stellen [12]

We kunnen data op verschillende manieren ter beschikking stellen aan de onderneming. We kunnen dit doen door gebruik van een traditioneel datawarehouse dat via ETL-systemen geladen gaat worden. In recente gevallen gaan er nu data lakes ontstaan of aangelegd worden. Deze bevatten alle ruwe data die een bedrijf kan gebruiken om analyses op uit te voeren. Door big data is het mogelijk om veel data goedkoop op te slaan.
Door gebruik te maken van een 'sandbox' (een speelterrein) kunnen de data scientists al bepalen welke informatie interessant is. Door gebruik te maken van een ETL-tool gaan ze vanuit de data lake de juiste gegevens ophalen en deze naar het sterdiagram op het dataware house wegschrijven voor operationeel gebruik.

2.3 Best practices

2.3.1 Best practices voor analytics [13]

Best practices zijn methoden die algemeen aangenomen worden als het meest efficiënt en vaak beter zijn als de standaard manier om iets te doen. Tijdens software development is het een methode die bijdraagt tot een succesvol onderdeel van het development proces. Het gebruik van best practices is nuttig voor bepaalde taken in het bedrijf die niet of moeilijk te standaardiseren zijn.

---

[11] J. VAN OVERVELDT, *3_DataGovernance*, Antwerpen, Karel De Grote-Hogeschool, l.d.,16 – 21; MICROSOFT, https://msdn.microsoft.com/en-us/library/bb190163.aspx, geraadpleegd op 23 januari 2017.
[12] J. VAN OVERVELDT, *5_DataBeschikbaarMaken*, Antwerpen, Karel De Grote-Hogeschool, l.d.,7-9.
[13] QUALITY BUSINESS SUPPORT BLOG, https://qualitybs.wordpress.com/2012/01/29/vastleggen-van-best-practices/, geraadpleegd op 24 januari 2017.

Dit kan zijn omdat deze taken heel hard afhangen van bepaalde omstandigheden. Om een voorbeeld te geven: er moet telkens een bepaalde taak uitgevoerd worden, bij deze taak moet je gaan kijken naar de situatie en hierop een beoordeling maken zodat het verloop van de taak aangepast wordt om efficiëntie te behouden. Best practices gaat hiervoor de vaststellingen van vorige iteraties van deze taak samenbrengen, en hiervoor een stappenplan opstellen. Dit stappenplan bevat dan hoe er het best gehandeld kan worden zodat er het meest efficiënt vooruitgang geboekt wordt.

Een paar kenmerken van best practices zijn makkelijk te identificeren.

1. Tijdsgebonden

Ze zijn afhankelijk van eerder opgedane kennis; ze kunnen niet zomaar opgesteld worden voor een taak die nog nooit is uitgevoerd. Ze hebben een ontwikkelingsproces nodig.

2. Toegevoegde waarde

De toegevoegde waarde van de best practice moet duidelijk bewezen zijn.

3. Voor alle situaties geldt:

De reden dat best practices gemaakt zijn is omdat het verloop van een taak afhangt van bepaalde omstandigheden. Daarom worden deze niet zo gedetailleerd gemaakt; ze geven een algemeen idee wat er moet gebeuren in welke situatie maar gaan nooit zeer specifiek zijn.

3. Literatuurstudie

Teneinde een bevattelijk corpus van best practices bij big data analytics samen te stellen, hebben we twee soorten bronnen geraadpleegd, met name: (1) literatuur (in de vorm van white papers, magazineartikels en artikels op websites) en (2) interviews afgenomen bij lokale experts. Wat hier volgt, is het resultaat van ons onderzoek op basis van die eerste soort bron, namelijk de gangbare literatuur.

Uit de geraadpleegde literatuur, hebben wij de volgende vier best practices gedistilleerd:

1. Hou de nodige investering zo klein mogelijk;
2. Begin klein en traag;
3. Gebruik 'clean data';
4. Onderschat het belang van 'sandboxes' niet.

3.1 Hou nodige investering zo beperkt mogelijk[14]

Big data analytics zijn enkel behulpzaam voor je bedrijf in zoverre de inzichten die eruit voortkomen, kunnen worden omgezet in een reële verbetering van de bedrijfsvoering. Daarom is het belangrijk om te weerstaan aan de neiging om alles zomaar te gaan analyseren. Focus altijd op wat écht belangrijk is. Bepaal wat voor analyses zinvol zouden zijn, voor je erin investeert.

Eén mogelijke manier om je ervan te verzekeren dat je analyseprojecten steeds relevant en verantwoordbaar zijn, is om een formeel proces uit te werken waarin de link duidelijk wordt tussen het resultaat van analyses en de impact die deze op de business heeft.

Het is ook steeds een goed idee om big data te analyseren in termen van relevantie voor gestructureerde data die je al hebt (bv. door de big data te filteren op informatie die aan je beste klanten gekoppeld kan worden). Een dergelijke aanpak verzekert relevantie en heeft vaak een significante impact.

Tools voor geavanceerde analyses dienen enkel gebruikt te worden indien dit gedicteerd wordt door de complexiteit van de onderzoeksvraag. Investeer enkel

[14] M. BARBERO, e.a., *Big data analytics for policy making*, European Union, 2016, 91-92; X, "5 best practices for big data analytics", Network World Asia, nov/dec 2015, 6; X, *An Enterprise Architect's Guide to Big Data*, Oracle Enterprise Architecture White Paper, maart 2016, 43-44.

in een geavanceerde analyse als het écht nodig is (en als er dus geen alternatieve methodes te bedenken zijn).

Uiteindelijk is big data analytics niet meer dan één manier om een bepaald doel te bereiken; andere methodes gebruiken, of deze combineren met big data analytics, kan zinvol zijn.

3.2 Begin klein[15]

Veel bedrijven zijn geïntimideerd door het concept van big data analytics en data driven organisations. Maar een best practice van big data analytics is dat je best klein begint, door analyses te maken over die niet lang duren over onderwerpen die weinig data nodig hebben.

In andere bedrijven heb je het probleem dat ze er te enthousiast aan beginnen. Hierdoor zijn er platformen ontwikkeld die geen goede infrastructuur bieden. Doordat de platformen niet optimaal ontwikkeld zijn en er veel geld in geïnvesteerd is, willen sommige bedrijven het platform laten herwerken. Er wordt dan terug veel geld geïnvesteerd en is er nog geen zekerheid dat het platform een goede infrastructuur bezit. Hierdoor gaan er aanzienlijke hoeveelheden geld verloren.

Ze zien het nogal snel te groot en gaan hierdoor onnodig veel kosten doen om bijvoorbeeld de infrastructuur te krijgen. Een beter idee is dan om te beginnen met cloud based data storage in kleine hoeveelheden. Want ondanks dat veel big data programs gratis en open source zijn, kunnen de kosten van big data analytics in een bedrijf zich snel opbouwen. Als er een groot analyse project wordt opgezet door een bedrijf dat weinig ervaring heeft, dan is de kans groot dat het fout gaat. Door de lange duurtijd van dit project gaat de trial en error van deze use case langer duren. De kostprijs van dat analyse project gaat dan veel groter zijn dan verwacht. Het belangrijkste is dus dat je klein begint en met kleine stappen meer ervaring opbouwt.

[15] BIGDATAWEEK, http://blog.bigdataweek.com/2016/02/23/keys-big-data-start-small-think-big-grow-fast/, geraadpleegd op 26 januari 2017; X, "5 best practices for big data analytics", Network World Asia, nov/dec 2015, 6.

3.3 Gebruik 'clean data'[16]

Veel bedrijven willen meevolgen met de nieuwe big data hype. Ze beginnen er vaak aan zonder een duidelijk plan of data management programma. Hierdoor gaan ze al hun big data ophalen en bijhouden zonder verfijning. De hoeveelheid ruwe data die ze dan bijhouden is vaak onnodig groot en substantiële delen ervan worden waarschijnlijk niet eens gebruikt. De reden dat bedrijven deze data dan toch bijhouden is voor het geval dat ze nog bruikbaar kunnen zijn bij toekomstige query's. Het is dus gewoon uit 'veiligheid' dat veel onnodige data wordt bijgehouden. Als een bedrijf ervoor kiest om data toch bij te houden moeten ze er ook wel een data scientist opzetten die in de ruwe data gaat zoeken naar zaken die een toegevoegde waarde kunnen creëren. Als ze dit niet doen dan is er geen meerwaarde om de data bij te houden om ooit eens te gaan gebruiken. (Dit wordt verder uitgelegd in het volgend punt, over de 'sandbox'.)

Het is dus belangrijk om niet te veel data bij te houden, maar je mag zeker ook niet te veel data strippen. Het beste is dat er wordt gekeken naar wat de meeste impact heeft. De kost van het bijhouden van data moet vergeleken worden met de kans dat die data in de toekomst nog gebruikt zou kunnen worden.

Door het ontwikkelen van bepaalde regels en methodes om big data te verwerken worden er zo veel kosten vermeden worden en gaat er geen belangrijke data verloren. Bij big data analytics draait tenslotte 70% van het werk rond het verzamelen van data en 30% rond het analyseren. Als het verzamelen van data niet goed gebeurt, kan de analyse ook geen uitgebreid en accuraat resultaat teruggeven.

3.4 Onderschat het belang van 'sandboxes' niet[17]

Een data lake heeft een interface die ook wel een sandbox genoemd wordt. Een sandbox is iets dat de data scientist kan gebruiken om te experimenteren en testen. Omdat het compleet gescheiden staat van de productieomgeving is dit een veilige omgeving om in te werken. Aan de hand van deze tool kan hij zien welke analyses een toegevoegde waarde kunnen opleveren. Als een data scientist iets gevonden heeft dat een toegevoegde waarde kan geven, gaat hij het via een ETL-tool naar het datawarehouse of operationele omgeving transformeren. Als dit gebeurd is, kunnen er effectief analyses op uitgevoerd

---

[16] INFORMATIONWEEK, http://www.informationweek.com/big-data/big-data-analytics/structuring-your-data-team-9-best-practices/d/d-id/1323601?image_number=6, geraadpleegd op 26 januari 2017.

[17] TECHOPEDIA, https://www.techopedia.com/definition/28966/data-sandbox-big-data, geraadpleegd op 26 januari 2017; X, *An Enterprise Architect's Guide to Big Data*, Oracle Enterprise Architecture White Paper, maart 2016, 44; M. BARBERO, e.a., *Big data analytics for policy making*, European Union, 2016, 95.

worden.

In de sandbox kan er ook onverwachte load getest worden. Het wordt ook vaak gebruikt om op data te gaan experimenteren. Het is de ideale omgeving voor data scientists omdat ze hier ongestoord gegevens uit ruwe data kunnen halen. Deze gegevens kunnen dan later gebruikt worden om een toegevoegde waarde te creëren.

4. Interviews

4.1 Situering

Dit gedeelte gaat over de interviews die wij zijn gaan afnemen bij verschillende bedrijven die gespecialiseerd zijn in big data en/of big data analytics. De bedrijven die we hiervoor gecontacteerd werden, zijn Deloitte, Big Industries, I4BI, Infofarm en SAS Institute.

De bedrijven Big Industries, I4BI en Infofarm zijn een onderdeel van de Cronos Groep. Van deze groep waren Big Industries en Infofarm voornamelijk gespecialiseerd in big data. Als uitbreiding bood Infofarm op maat gemaakte tools om analyses op uit te voeren. I4BI is een nieuwe speler op vlak van big data maar hebben hun ervaring reeds behaald in het analytics gedeelte.
Wij zijn bij SAS Institute Belgium langs geweest, dit is een onderdeel van SAS en zij zijn een vendor voor alle SAS-producten.
Deloitte is ook gespecialiseerd in big data maar ook in big data analytics.

Alle interviews zijn afgenomen met dezelfde vragen. Dit om zoveel mogelijk eenheid in de interviews te zien.

De vragen zijn doelbewust redelijk vaag gehouden, zodat de bedrijven hun eigen zicht op bepaalde onderwerpen konden geven. Zo konden we eerlijke antwoorden van verschillende bedrijven met elkaar vergelijken. Tevens konden we direct zien welke vragen een gelijkaardig antwoord kregen bij de geïnterviewde personen.

4.2 Doel

Het doel van deze interviews was om meer te weten te komen over big data en big data analytics in het effectieve bedrijfsleven. Hoe door ervaring hun kijk op de zaken was bepaald. Door welke ervaring welke best practices naar voren gekomen zijn. Dit willen we dan vergelijken met de puur theoretische voorbeelden die we online tegengekomen zijn. Ook konden we zo achterhalen hoe een bedrijf dat zich focust op deze onderwerpen zich gedraagt, open staat voor mensen die vragen komen stellen etc.

4.3 Voordat we beginnen

In deze sectie gaan we de interviews bespreken in termen van overeenkomsten en verschillen. In bijlage zijn beknopte schriftelijke verslagen opgenomen van alle interviews (inclusief vragen, gegroepeerd per bedrijven). Audio-opnames

van de interviews zijn in ons bezit, maar worden niet publiekelijk ter beschikking gesteld om privacy-redenen. Op aanvraag van een docent, in het kader van de beoordeling van onze paper, kunnen deze opnames wel worden geconsulteerd.

4.4 Wie hebben we geïnterviewd

Wij gaan eerst alle personen voorstellen die we hebben geïnterviewd. Zo kan een duidelijk beeld ontstaan over hun competenties en ervaringen.

4.4.1 SAS Institute[18]

Bij SAS hebben wij het genoegen gehad om drie personen te interviewen.

De eerste persoon was Elisabeth Versailles. Zij werkt ondertussen al 21 jaar voor SAS. Ze heeft hier verschillende functies gehad waaronder lesgeven in SAS. Nu is ze voornamelijk bezig als Academics Relationship Manager. Dit wil zeggen dat zij de contacten verzorgt met alle academische instituten en hun cursisten. Zij was onze contactpersoon om alles bij SAS Institute te regelen. Als studies heeft zij een Master in geschiedenis gedaan. Sinds 1996 is ze begonnen bij SAS en is ze doorgegroeid van trainingen geven naar lesgeven in 2003 tot interne projectmanager. Dit heeft ze 2 jaar gedaan en is uiteindelijk in 2013 tot haar huidige functie bekomen.

De tweede persoon was Tine Van Dyck. Zij heeft ondertussen al rond de 10 jaar ervaring als Business Intelligence analist. Zij tekent voornamelijk de strategieën uit. Ze gaat analyseren hoe een klant vandaag de dag werkt en gaat dan een strategie uit teken hoe zij dit op een gestructureerde manier kunnen laten verbeteren. Zij heeft een master in de wiskunde behaald met optie beleidsinformatica. Ze heeft een lange tijd gewerkt als functioneel analist bij een bank. Hier heeft ze voornamelijk het Datawarehouse opgezet. Dit heeft ze 4 jaar gedaan en is dan bij SAS terechtgekomen. Na 5 jaar algemene sales support te zijn geweest is ze doorgegroeid tot big data (analytics) customer business manager.

Onze derde persoon was Anthony Severeyns. Hij heeft ondertussen al 11 jaar ervaring met Business Intelligence analyses. Waaronder 8 jaar ervaring bij SAS Institute zelf. Hier is hij nu pre-sales consultant. Dit betekent dat hij voornamelijk bezig is met naar klanten of prospecten toe gaan. Hier gaat hij voornamelijk demo's geven over de producten van SAS en hoe ze de klanten optimaal kunnen helpen. Ze gaan hier bepalen wat er relevant voor de klant is en welke waarden zij er kunnen uithalen. Hij is meer de persoon voor de juiste

[18] E. Versailles, *Mondelinge mededeling* (interview), Tervuren, 18 januari 2016, 13.30 uur; T. Van Dyck, *Mondelinge mededeling* (interview), Tervuren, 18 januari 2016, 13.30 uur; A. Severeyns, *Mondelinge mededeling* (interview), Tervuren, 18 januari 2016, 13.30 uur.
bijlage 10.1.5

technologie te bepalen, zoals het data management SAS-platform. Hij heeft burgerlijk ingenieur elektronica gestudeerd. Hierna is hij doorgegaan voor een master in de bedrijfseconomie. Zijn eerste werkervaringen waren bij het bedrijf Materialize. Hier heeft hij voornamelijk aan prototyping van medische implantaten gewerkt. Hierna heeft hij 3 jaar gewerkt als Business Intelligence, budget en planning consultant.

4.4.2 Cronos - Big Industries [19]

Bij Big Industries hebben we het genoegen gehad om één van de medeoprichters te spreken. Namelijk Matthias Vallaey.

*Figuur SEQ Figuur * ARABIC 5 Logo Big Industries*

Hij heeft ondertussen al 4 jaar ervaring in Big Data. Hij heeft een master in de toegepaste economische wetenschappen behaald. Verder heeft hij al reeds 23 jaar werkervaring opgedaan en dit voornamelijk in de IT bij verschillende vendors. Hij is dan 4 jaar geleden begonnen bij Cronos vanuit een ondernemerschap. Hij heeft toen gekozen om een bedrijf rond Big Data op te starten vanwege dit toen echt nog een hype maar niemand echt wist hoe je hieraan moest beginnen. Hij is bij Cronos uitgekomen omdat dit een typisch bedrijf is dat op elke nieuwe technologie wil inzetten en mee zijn met zijn tijd. In de begintijd zijn ze voornamelijk gaan analyseren hoe je een big data project opbouwt. Dit hebben ze gedaan door trainingen te gaan volgen en gericht te gaan werken. 3 jaar geleden is hun eerste klant (Telenet) bij hen gekomen. Na een evaluatie met Cronos hebben ze besloten dat Big Data een levensvatbaar begrip en ideologie was en hebben ze besloten om Big Industries in het leven te roepen.

4.4.3 Cronos - Infofarm [20]

*Figuur SEQ Figuur * ARABIC 6 logo Infofarm*

Bij Infofarm hebben we het genoegen gehad om Ben Vermeersch te mogen interviewen. Hij is voormalige KdG-student Toegepaste Informatica met als afstudeerrichting Applicatieontwikkeling. Hij is bij Cronos begonnen en heeft ons iets meer uitleg geven over de Cronos groep. Hier zijn we te weten gekomen dat de Cronos groep een 300 tal bedrijven heeft met ongeveer 4 tot 5 duizend werknemers. Elk bedrijf focust op zijn eigen technologieën maar hebben wel een gemeenschappelijke kern (Cronos). Er zijn ook veel subgroepen waaronder Xplore Group, Crosspoint Solutions, …

Hij is in 2014 begonnen vanuit een technisch Java idee met Infofarm. Maar ze

---

[19] M. Vallaey, *Mondelinge mededeling* (interview), Kontich, 17 januari 2016, 13.30 uur. bijlage 10.1.2

[20] B. Vermeersch, *Mondelinge mededeling* (interview), Kontich, 17 januari 2016, 10.00 uur.
bijlage 10.1.3

wilden meer met big data gaan werken en iets met de combinatie tussen big data en data science. Hierdoor is in 2015 Infofarm een zelfstandig onderdeel geworden onder Cronos - Xplore Group.

4.4.4 Cronos - I4BI [21]

Bij I4BI hebben we het genoegen gehad om met Hannes Cassiers en Maarten Van den Broeck te kunnen spreken.

Hannes Cassiers heeft ondertussen al 13 jaar ervaring als datawarehouse specialist en business intelligence analist. Hij heeft ook 2 jaar ervaring als Big Data specialist. Hij heeft weinig theoretische achtergrond rond big data en meer ervaring rond de opvolging van wat er leeft en wanneer je aan big data moet beginnen. Ze hebben ook de eerste big data stage vorig jaar uitgestuurd.

Hij is afgestudeerd als industrieel ingenieur elektronica met optie ICT. Dit was in de Karel de Grote Hogeschool campus Hoboken. Hierna is hij gestart bij een consultancy firma en heeft hij genoten van een opleidingen in transactionele systemen. Hierna heeft hij een volgend project gekregen dat een business intelligence project was. Dit heeft hij 10 jaar gedaan. Sinds 2016 is hij begonnen bij XploData.

Maarten Van den Broeck is een buitenbeentje in de big data wereld. Dit komt omdat hij recent begonnen is, namelijk in november 2016, en heeft een doctoraat in de biologie behaald. Hij heeft hiervoor een thesis geschreven dat aanleunt bij een hackathon. Hij was hier wel benieuwd in en is naar de hackathon geweest. Dit heeft zijn interesse in big data en data scientist opgewekt en is zo begonnen bij I4BI. Hij heeft dus tot op heden nog maar 2 maand ervaring.

[21] H. Cassiers, *Mondelinge mededeling* (interview), Kontich, 18 januari 2016, 10.00 uur; M. Van den Broeck, *Mondelinge mededeling* (interview), Kontich, 17 januari 2016, 13.30 uur.
bijlage 10.1.1

4.4.5 Deloitte [22]

Onze contactpersoon bij Deloitte was Els Renders. Wegens omstandigheden hebben wij haar niet kunnen ontmoeten of rechtstreeks kunnen aanspreken. Ze heeft ons geholpen door een schriftelijk interview af te nemen.

Ze heeft reeds 20 jaar ervaring als business intelligence analist. Sinds 2012 is ze in het big data verhaal gestapt. Hierdoor heeft ze ongeveer 5 jaar ervaring in big data. Ze heeft een diploma in arbeidssociologie met een optie survey analytics behaald. Ze heeft al heel veel werkervaring waaronder survey bij Dimarso, Analytics bij de Federale politie, beheer van gestructureerde en ongestructureerde data en business intelligence bij Flanders Investment & Trade, Business intelligence consultant bij Numius en uiteindelijk aangeworven geworden door Deloitte. Hier is ze nu Analytics & Information Management Consultant.

4.5 Wat is er gezegd geweest

Uit de interviews bleek dat er verschillende mogelijke profielen mogelijk zijn. We zien twee grote groepen om ze in te verdelen. Namelijk het technisch aspect waarin we de technologieën bedoelen. Het andere is het eerder praktische aspect zoals een data scientist of een citizen data scientist. Er wordt niet echt vastgehouden aan de 2 profielen want alles ertussen komt ook voor. Er is zeker nog een tekort aan mensen die zich bezighouden met analytics. Dit wordt veroorzaakt doordat het zeer moeilijk is om de juiste profielen te vinden. Veel bedrijven weten ook niet hoe ze deze profielen moeten gaan identificeren. Je vindt namelijk heel moeilijk complete data scientists. Hiermee bedoelen we data scientists die op elk vlak excelleren. Ook zijn er weinig die effectief als data scientist te werk gaan. Ze proberen dit op te lossen door enerzijds teams samen te stellen om zo aan de juiste profielen te geraken. Anderzijds wordt het tekort aan analisten opgelost door mensen van development die zich omscholen. Door de omscholing is het concept van Citizen Data Scientist ontstaan. Hoewel analytics als knelpuntberoep wordt bezien en de vraag nog zeker gaat toenemen over de komende jaren, kunnen we besluiten dat er zeker in andere onderdelen van IT, zoals .NET, evenveel vraag is. Om het juiste profiel te vinden ga je niet moeten zoeken in de business of technische kant alleen maar heb je eerder een mix nodig. Analytics is namelijk het kruispunt waar dat de business de IT raakt. Bij net afgestudeerden moet er interesse zijn in het andere aspect om een ideale toekomstig profiel te onderbouwen. Een belangrijke skill is voornamelijk communicatie. Een data scientist kan alleen werken maar als deze de communicatie skill, of storytelling, mist kunnen er veel fantastische modellen zijn die niet verkocht geraken aan de business. Waardoor deze ook niet zullen gebruikt worden. De grootste uitdaging is een balans vinden tussen al deze aspecten.

---

[22] E. Renders, *Mondelinge mededeling* (bijlage in mail), Kontich, 15 januari 2016, 13.30 uur.
bijlage 10.1.4

Hieruit concluderen wij onze eerste best practice, namelijk de juiste profielen aanwerven voor de juiste taken. Het maakt niet uit of ze uit Business of IT komen. Ze moeten interesse hebben in het andere aspect.

We hebben in onze interview gepolst naar de 3 belangrijkste aspecten die een data scientist moet hebben. Het meest voorkomende antwoord was namelijk statistiek. Hiermee wordt bedoeld de affiniteit hebben met cijfers en data. Je moet kunnen bepalen of uw bronnen juist zijn. Maar ook hiermee kunnen omgaan. Het tweede belangrijkste aspect dat werd aangehaald is business kennis maar ook computer skills. Hiermee wordt bedoeld dat je de bedrijfskennis moet hebben om bepaalde zaken juist te achterhalen. Als je helemaal geen basiskennis hebt van hoe bijvoorbeeld de riolering werkt, moet je ook geen analyses hierop gaan uitvoeren. Want dit kan resulteren in foutieve voorspellingen. Als je geen computer skills hebt ga je ook de data niet kunnen verwerken op een correcte manier. Ten slotte kwam communicatie skills altijd mee naar boven. Je kan nog een goochelaar zijn met data en cijfers, je kan de beste modellen maken. Maar kan je ze niet uitleggen en verklaren aan de business dan gaan zij hier geen meerwaarde in kunnen zien. Om nog aan te vullen hebben we een bonus aspect namelijk leergierigheid Als je leergierig gaat zijn ga je veel meer kunnen bereiken in het analytische wereldje.
Het is ook belangrijk om de juiste hoeveelheid data te hebben. Als we kiezen voor te weinig data kan dit een vertekend beeld geven voor de use case.
Het is ook aan te raden om alle ruwe data bij te houden in een data dump. Dit verschilt tegenover een data lake. Een data lake heeft nog een interface bovenop de data staan, een data dump niet.

Hieruit concluderen we onze tweede best practice, namelijk de juiste bronnen kiezen en uw data valideren. Als je foutieve bronnen gaat gebruiken kun je nog zoveel doen als je wil, maar gaat het eindresultaat nooit zo goed zijn.

We hebben ook eens gepolst of onze geïnterviewde personen big data nog als een hype zagen. 1 bedrijf ziet big data nog steeds als een hype en zegt dat alle bedrijven moeten meedoen of ze gaan ten onder. Maar hij geeft wel toe dat er veel te veel bedrijven hier gewoon opgesprongen zijn zonder te weten wat ze doen. Alle andere bedrijven vonden voornamelijk dat big data over de hype heen was en er nu gewoon is. Dit wordt ook onderbouwd door de Gartner Hype Cycle. We moeten nu gewoon met data omgaan of deze nu big data of small data is. Tijdens de hype zijn er heel veel bedrijven die platformen voor big data online hebben gezet die niet aansloegen bij het doelpubliek. Hierdoor is er heel veel geld verloren gegaan. Heel veel bedrijven hebben big data als hype erkent en zijn hier zomaar op gesprongen zonder te weten wat ze moesten doen. Hoewel big data over zijn hype heen is zijn ze wel algemeen eens dat het nog steeds veel beloven is. We kunnen ook concluderen dat België nu begint mee te gaan met het big data verhaal, dit kan je zien aan bijvoorbeeld de mediahuizen en telecombedrijven. Big data en Data Analytics zijn voornamelijk buzzwords geweest die mensen hebben aangezet tot het investeren in technologieën zoals

een hadoop cluster. Dit was niet genoeg. Het belangrijke gedeelte kwam hierna pas en dat was inzichten halen uit de vergaarde data die men had verzameld en de mogelijke manier om deze zinvol te verwerken.

Hieruit concluderen wij onze derde best practice, namelijk dat je niet zomaar op een hype moet springen zonder dat je weet wat je doet. Begin klein en zorg dat je zo snel mogelijk faalt om hieruit te leren. Door alles zomaar te volgen zonder te weten wat je doet, ga je veel geld en tijd verliezen.

We hebben ook gepolst naar welke platformen (big data en analytics) ze het meeste gebruiken. Als big data kwam het hadoop ecosysteem altijd terug. We denken hierbij aan de 2 grootste spelers Cloudera en Hortonworks. Deze zijn cruciaal om big data projecten mee door te voeren. Een onderdeel van hadoop is Spark dit maakt dat het een ideaal ecosysteem is om applicaties mee te maken en uit te voeren. Veel bedrijven willen ook niet meer on premise gaan werken en kiezen daarom voor een cloud oplossing. De grootste en belangrijkste spelers hiervoor zijn Amazon Web Services en IBM. In bepaalde gevallen kiezen de bedrijven er wel voor om on premise te werken maar hebben ze geen kennis of geld om zelf een architectuur op te zetten. Daarom werken ze samen met Oracle. Ze nemen hier een Oracle Big Data Appliance. Dit is een server dat ze komen leveren en hierop draait een instantie naar keuze, het meest voorkomende is Cloudera. Als je kiest voor een vendor kun je ook kiezen voor SAS data processing. In de meeste gevallen wordt er een data lake aangelegd. Om hier nu een toegevoegde waarde op te krijgen ga je analyses moeten uitvoeren. Hiervoor kan je opteren voor een Aster DB. Dit is een Databank die verschillende soorten databanken kan gaan connecteren en gaat voor jou hier dan analyses op uitvoeren. Het nadeel hieraan is dat je niet weet hoe dit werkt. Als je zelf in de hand wil houden hoe je analyses gaat uitvoeren en hoe het achterliggend werkt ga je andere tools gebruiken. Deze tools kunnen SAS Visual Analyzer, Oracle Big Data Discovery Tool, Oracle BI, Teradata zijn. Je kan ook opteren om analyses met R uit te voeren. Als je volledige controle wil over het analyse gedeelte is het aan te raden om zelf een tool te schrijven.
Maar het platform gaat zelden de motor van het succes zijn of de bron van het falen. Je moet altijd weten wat de mogelijkheden of de beperkingen zijn van jouw platform. Transparantie naar de betrokken partijen is hier een must. Dit samen met openheid zorgt voor een optimale werking. De use case bepaalt de tool. Je gaat op basis van de use case de optimale tools keuzes moeten maken.

Hieruit concluderen wij onze vierde best practice, namelijk de use case is de bepalende factor van de te gebruiken tools en platformen. Als je een bepaalde tool kiest waarbij de use case helemaal niet tot zijn recht komt, kan dit leiden tot het falen van een project.

Uit de interviews is gebleken dat Master Data Management heel cruciaal is om de data af te schermen voor de eindgebruiker. Dit heeft al voordeel dat er security en privacy laag optreedt. Het aanleggen van meta data en meta data dictionary helpt de gebruikers ook te identificeren welke begrippen bij wat horen.

Hieruit concluderen wij onze vijfde best practice, namelijk het gebruik van meta data dictionary en master data management. Dit om de gebruiker het makkelijk te maken maar toch hun van het bronsysteem weg te houden.

Het belangrijkste voor een bedrijf dat wil starten met Big Data moet op de volgende zaken letten. Het belangrijkste waar een bedrijf op moet focussen is innovatie. Als ze niet innoveren dan is er meer kans dat dit het einde betekent. Dit kunnen ze gaan verwezenlijken door hun mensen te gaan laten bijscholen in de verschillende competenties en technologieën. Dit creëert een toegevoegde waarde. Er moet dialoog geweest zijn tussen Business en IT om de zinvolle toepassingen te selecteren. Hierna kan er best zo snel mogelijk 'trial and error'-projecten opgestart worden. Door de verschillende aanpakken te gaan testen kan er zo ervaring opgedaan worden voor toekomstige projecten. Door de juiste profielen te kiezen en de beste use case te gaan uitwerken gaan we de toegevoegde waarde effectief creëren. Hierna kan je ook effectief nagaan of je de juiste toepassingen geselecteerd hebt. Je moet ook voor jezelf gaan uitmaken of je bestaande tools gaat gebruiken of dat je gaat investeren in een eigen tool te gaan maken. Hiervoor treedt wel het probleem op dat je technische kennis nodig hebt. Maar dit kan je oplossen door de juiste profielen aan te trekken of mensen te laten bijscholen. Je moet ook verder kijken dan de ruwe data alleen. Als er al een datawarehouse is, moet je proberen deze ook te gaan gebruiken. Je moet dus rekening houden met wat er al reeds bestaat.

Dit alles kan leiden tot kostenbesparing, een toegevoegde waarde voor de business creëren, zaken die moeilijk waren voor de komst van Big Data zijn ineens veel makkelijker vb. data opzoeken. Dit kunnen we allemaal bereiken door de processen te optimaliseren. Dit gaat ook nog als gevolg hebben dat we relevante patronen gaan kunnen opsporen waardoor we meer inzicht krijgen in onze onderneming. We gaan onze beslissingen laten automatiseren op basis van modellen. Hierdoor kunnen we ons buikgevoel ontkrachten of bevestigen.

Hieruit concluderen wij onze zesde best practice, namelijk trial and error. Door zoveel mogelijk aanpak te gaan testen kan je ervaring opdoen en snel beslissen welke aanpak er juist is en welke niet.

We hebben ook gepolst naar waar je op zou moeten letten als je een analytics platform opstelt. We zijn hierbij achter komen dat schaalbaarheid en flexibiliteit heel belangrijk zijn. We moeten het platform kunnen aanpassen naar gelang de capaciteit het nodig heeft. IT en business moeten dit platform kunnen gebruiken. Het mag dus niet alleen maar IT gericht zijn. Een oplossing kan in theorie heel makkelijk en de beste lijken, maar kan in de praktijk heel moeilijk tot stand komen. Je moet soms zelf specialisten gaan raadplegen. Je moet openstaan voor andere oplossingen om efficiënt te werk blijven gaan. Het is ook belangrijk om te bepalen wie je doelpubliek is zodat je de juiste keuzes kunt maken van wat er in het platform moet steken of hoe gebruiksvriendelijk het platform moet zijn, voorbeeld een business persoon houdt misschien meer van een point-click interface, terwijl een IT gerelateerde persoon meer van een command gebaseerde interface opteert.

Hieruit concluderen wij onze zevende best practice, namelijk openstaan voor andere oplossingen. Je kan zo goed en zo kwaad als je maar wil achter je oplossingen blijven staan. Maar als een andere oplossingen voordeliger of beter is, moet je dit kunnen accepteren.

We concluderen hier ook nog onze achtste best practice, namelijk doelpubliek bepalen. Je kan een platform opstellen die te ingewikkeld of te veel omvat. De eindgebruikers gaan dit dan te veel vinden en het niet meer gebruiken.

Het doel van deze interviews was om een roadmap te bekomen om van A tot Z een platform op te stellen. Toen we deze vraag stelden kwamen we tot deze antwoorden. Het belangrijkste is dat je klein moet beginnen. Je moet use cases kiezen waarvan je weet dat je ze van end-to-end kunt uitwerken. Het is ook belangrijk om relevante use cases te kiezen en deze te prioriteren. Je moet kritisch staan tegenover de noden van de klant of eindgebruikers. Je moet bepalen wat er haalbaar is en wat niet. De bedoeling is voor kostenefficiënt te werken en niet om nog meer te gaan besteden aan projecten die zeker gaan falen. Je moet de klant niet meer geven dan wat hij vraagt want dit is een nefast voor de gebruiksvriendelijkheid van de klant en voor het kostenefficiëntie plaatje. Je moet op basis van de use cases gaan bepalen wat de juiste tools zijn en deze gaan gebruiken. Je kan ook best rekening houden met de kwaliteit en de betrouwbaarheid van de bronnen en bronsystemen. Je moet altijd in je achterhoofd houden dat de combinatie met strategie, profielen, technologieën data management en processen de sleutels zijn tot een geslaagd project. Een belangrijk aspect is ook de overtuigingskracht, via visualisatie, om de ondernemingen te overtuigen dat er change management nodig is om kostenefficiënt te gaan werken. Om een sterke tool te creëren moet je de juiste technologie kiezen en deze aanvullend met de menselijke input.

We concluderen hieruit onze negende best practice, namelijk klein beginnen. Je moet een use case kiezen waarvan je weet dat hij end-to-end afgewerkt kan worden. Het is beter om een klein project volledig af te krijgen, dan een groot project niet af te krijgen.

We concluderen hieruit ook onze tiende best practice, namelijk overtuiging. Er is overtuigingskracht nodig om de veranderingen verkocht te krijgen aan de onderneming. Je moet de onderneming doen inzien dat investeren in een big data/big data analytics platform tot kostenefficiënt werken kan leidt.

We concluderen hieruit ook onze elfde best practice, namelijk menselijke input. Om tot een ijzersterke tool te komen moet je de juiste technologieën gebruiken en deze koppelen aan de menselijke input.

Doorheen de interviews hebben we ook nog de volgende best practices tegengekomen die we niet hebben gegroepeerd bij andere onderdelen.

We hebben onze twaalfde best practice geconcludeerd, namelijk durven. Je moet durven uit je comfortzone te komen en durven ervoor te gaan.

We hebben onze dertiende best practice geconcludeerd, namelijk weten wat er leeft. Je moet meegaan met de technologieën en geïnteresseerd zijn om de nieuwe zaken te ontdekken.

We hebben onze veertiende best practice geconcludeerd, namelijk bekijk big data als een geheel. Je moet big data bezien als een geheel en niet een stukje van een bepaalde context.

We hebben onze vijftiende best practice geconcludeerd, namelijk samenwerking. Het is cruciaal dat er samenwerking optreedt tussen bepaalde profielen maar ook tussen bepaalde tools.

We hebben onze zestiende best practice geconcludeerd, namelijk opleiding. Je moet constant bijleren om mee te zijn met de nieuwste technologieën.

Hier volgen nog een paar voorbeelden van use cases die ze ons hebben meegedeeld. Om ze te groeperen gaan we het eerst hebben over de big data use cases. Hierin kwam heel veel data warehouse aan bod. Hier spreken we over het opzetten van nieuwe data warehouse met data lakes, bestaande vervangen tot zelf uitbreiden van bestaande. Het integreren van bestaande architecturen met big data architectuur of het importeren van big data. Het verplaatsen van business naar het data scientist verhaal hoort hier ook nog bij het big data verhaal, dit geldt ook voor ETL-tools en operationele support.

Bij analytics gedeelte komen we bij nog een paar meer use cases. Namelijk predictive of prescriptive analyses. Hier vinden we bijvoorbeeld onderhoudssystemen, life science research, impact beslissingen van een minister, uitbraak van menselijke virussen en op basis van stream analyses van het verkeer kunnen we bepalen of er file optreedt of niet. Een andere use case kan ook customer 360 zijn. Hierin vinden we turn prediction, in kaart brengen van wat de klant allemaal doet, verkoopcijfers en retail, IoT-producten. We hebben ook nog standaard rapporteringen om de ROI te verhogen, data processing toe te passen, Oracle BI Rapporteringen, Tableau en bedrijven te begeleiden om betere analyses uit te voeren. Overige use cases zijn nog security (intrusion detection), fraude (fraude detectie, fraude security en belastinginspectie/carrousels), nummerplaatherkenning, bouwen van een BI-oplossing, sensor data, clickstreams analyseren, segmentatie, clustering en optimalisatie.

5. Corpus van best practices

In de voorgaande twee delen hebben we een twintigtal best practices bijeengesprokkeld. Bedoeling is nu om deze te verwerken tot een bevattelijk corpus. Hiervoor nemen we de volgende drie stappen: (1) eerst maken we twee lijsten van alle verzamelde best practices, namelijk enerzijds deze die we uit de literatuur hebben gehaald, anderzijds diegene die we op basis van de interviews hebben geformuleerd; (2) vervolgens trachten we beide lijsten te combineren; (3) ten slotte structureren we de gecombineerde lijst tot een bevattelijker geheel, waardoor de samenhang van de best practices duidelijk wordt.

5.1 Oplijsting

In een eerste beweging, lijsten we alle verzamelde best practices op die we hebben verzameld.

5.1.1 Literatuur

Vanuit het bestuderen van de gangbare literatuur, hebben we de volgende lijst van best practices opgesteld:

1. Hou de nodige investering zo klein mogelijk
2. Begin klein
3. Gebruik 'clean data'
4. Onderschat het belang van 'sandboxes' niet

5.1.2 Interviews

Op basis van de interviews met lokale experts, hebben we de volgende best practices opgelijst:

1. Werf de juiste profielen aan voor de juiste taken
2. Kies de juiste bronnen en valideer je data
3. Begin klein en faal snel (bij het adopteren van een nieuwe hype)
4. Gebruik de use case om tools en platformen te bepalen
5. Gebruik een meta dictionary en master data management

6. Experimenteer zoveel mogelijk m.b.v. ‘trial and error’-projecten
7. Sta open voor andere oplossingen
8. Bepaal je doelpubliek
9. Begin met een kleine, relevante use case
10. Overtuig je onderneming van de noodzaak van big data analytics
11. Koppel menselijke input aan de juiste technologieën
12. Durf uit je comfort zone te komen
13. Blijf op de hoogte van wat er leeft in je kennisveld
14. Bekijk big data als een geheel
15. Zet samenwerking op tussen profielen en tools
16. Zet je opleiding continu verder

5.2 Samenvoeging

Vervolgens kijken we hoe we deze lijsten kunnen samenvoegen, en of hierbij enige conflicten ontstaan.

Wat we opmerken, is dat alle best practices die we uit de literatuurstudie hebben gehaald, terugkomen in de lijst van best practices die we hebben opgesteld op basis van de interviews.

De eerste best practice die voortkwam uit het bestuderen van literatuur, is “hou de investering zo klein mogelijk”. Onder deze noemer zitten een aantal concretere tips vervat (zie het corresponderende deel van deze paper), zoals “analyseer enkel relevante use cases”, “gebruik enkel een geavanceerde tool uit als de use case het vereist” en “overweeg alternatieven”, die stuk voor stuk terugkomen in de lijst van best practices uit de interviews, respectievelijk “begin met een kleine, relevante use case” (best practice 9), “gebruik de use case om tools en platformen te bepalen” (best practice 4) en “sta open voor andere oplossingen” (best practice 7).

De tweede best practice die we formuleerden op basis van onze literatuurstudie, was “begin klein”, wat deels overeenstemt met de derde best practice uit onze interviews (“begin klein en faal snel (bij het adopteren van een nieuwe hype)”) en deels met de negende (“begin met een kleine, relevante use case”).

Ten derde zien we dat de best practice uit de literatuur “gebruik ‘clean data’” overeenstemt met “kies de juiste bronnen en valideer je data”, de tweede best practice uit de interviews.

De vierde en laatste best practice die we uit de literatuur haalden, “onderschat het belang van ‘sandboxes’ niet”, stemt inhoudelijk overeen met “experimenteer zoveel mogelijk m.b.v. ‘trial and error’-projecten”, de zesde best practice die we uit de interviews hebben gedistilleerd.

Kortom, de samenvoeging van de twee bekomen lijsten met best practices, kan zonder conflicten gebeuren, en wel door simpelweg enkel de tweede lijst te behouden, aangezien de best practices van de eerste lijst toch mee in de tweede vervat zitten.

Merk ook op dat het voorkomen van bepaalde best practices in beide lijsten een significante bevestiging van deze best practices inhoudt.

5.3 Structurering

Onze derde en laatste stap in de constructie van een overzichtelijk corpus van best practices bij big data analytics op basis van de hiervoor verzamelde informatie, bestaat erin om de lijst van best practices die we hebben opgesteld, om te vormen tot een meer gestructureerd en bevattelijk geheel.

Concreter zien we de mogelijkheid om onze lijst van best practices in twee modellen te gieten: (1) de levenscyclus die een samenhangend geheel van best practices bij big data analytics in je organisatie zou kunnen volgen; (2) het trechtermodel, dat zou kunnen dienen als hulpmiddel om keuzemogelijkheden te filteren bij het nemen van beslissingen doorheen dit proces.

Een belangrijke opmerking bij deze modellen, is dat ze met een korrel zou dienen te worden genomen. Deze modellen dienen in de eerste plaats als een manier om de best practices op een meer gestructureerde manier voor te stellen. De manier waarop de verschillende best practices in deze modellen samenhangen, is een product van ons eigen denkwerk en is niet gebaseerd op bronnenonderzoek. Wat telt, zijn de individuele best practices. Wij hebben met deze modellen wel een poging gewaagd om de samenhang ertussen uit te tekenen, maar wij zijn zelf geen experts. Wat voor samenhang deze best practices in de realiteit hebben en hoe deze samenhang zou kunnen worden aangewend om reële bedrijfsprocessen rond big data analytics te optimaliseren, zou een interessante vraag kunnen zijn om aan experts in het domein te stellen, en zou derhalve een geschikte onderzoeksvraag kunnen zijn voor een vervolgonderzoek.

Desalniettemin dient het kind hier niet met het badwater te worden weggegooid; de modellen die we hebben geconstrueerd, vertonen een logische samenhang die de helderheid van de voorstelling van deze best practices ten goede komt.

5.3.1 Levenscyclus

Het merendeel van de best practices die we hebben geformuleerd, kan worden omgevormd tot een min of meer cyclisch geheel. Mogelijk stemt deze cyclus tot op zekere hoogte overeen met de het proces rond big data analytics dat wordt gevolgd in een organisatie die deze best practices hanteert.

1. Hype

   Als er een hype ontstaan moet je niet direct mee opspringen maar bepalen of dit een relevante hype is.

2. Durven

   Wanneer je bepaald hebt dat het een relevante hype is moet je durven uit je comfortzone komen en hierop springen.

3. Opleiding

   Wanneer je besloten hebt dat je op de hype springt moet je jezelf gaan opleiden in deze technologie.

4. Aanwerving

   Je gaat de juiste profielen moeten aantrekken om je project volledig te doen slagen.

5. Big data is een geheel

   Je moet Big Data niet bekijken als een klein blokje van analytics maar als het volledige geheel.

6. Klein beginnen

   Je moet een kleine use case gaan kiezen om hier mee te beginnen.

7. Bronnen kiezen

   Je moet de juiste bronnen gaan kiezen die je use case gaat ondersteunen en die relevant zijn.

8. Trial and error

   Je moet proberen zo snel mogelijk technologieën gaan uitsluiten door zo snel mogelijk met een bepaalde technologie te falen.

9. Openstaan voor andere oplossingen

Dit is een onderdeel van Trial and error. Wanneer we niet de juiste technologie hebben, moeten we openstaan voor een andere oplossing.

10. Use case bepalen

Wanneer we ervaring hebben opgedaan met de Trial and error kunnen we nu zelf use cases gaan bepalen die we effectief gaan uitbrengen.

11. Doelpubliek bepalen

We gaan bepalen voor welk doelpubliek onze use case relevant is en we gaan op deze basis beslissen welke technologieën we effectief gaan gebruiken.

12. Overtuigen

We gaan proberen onze doelpubliek warm te maken voor ons project, zodat ze hier mee willen instappen.

13. Meta data dictionary opstellen

Wanneer we de use case volledig gaan uitwerken, gaan we een meta data dictionary aanmaken om de eindgebruiker te informeren naar bepaalde termen die gebruikt zijn.

Opmerkingen:

Stappen 3 en 4 kunnen door elkaar gebruikt worden.

Stappen 8 en 9 kunnen meerdere keren achter elkaar herhaald worden.

5.3.2 Trechtermodel

De drie best practices uit onze lijst die nu nog overblijven, kunnen worden aangewend om keuzemogelijkheden te filteren bij het nemen van beslissingen inzake big data analytics, bijvoorbeeld inzake het selecteren van een bepaalde tool of technologie.

Specifieker kunnen, bij het overwegen van keuzemogelijkheden, drie stappen worden ondernomen om het aantal keuzemogelijkheden dat in overweging wordt genomen, te verkleinen. We passen met andere woorden een trechter toe om een overweldigend aantal keuzemogelijkheden te filteren tot een beheersbare, overzichtelijke hoeveelheid.

In eerste instantie, weet je als analist wat er leeft in de wereld rond technologieën voor big data analytics. Je blijft steeds goed op de hoogte van welke nieuwe technologieën recent zijn uitgekomen en probeert continu te

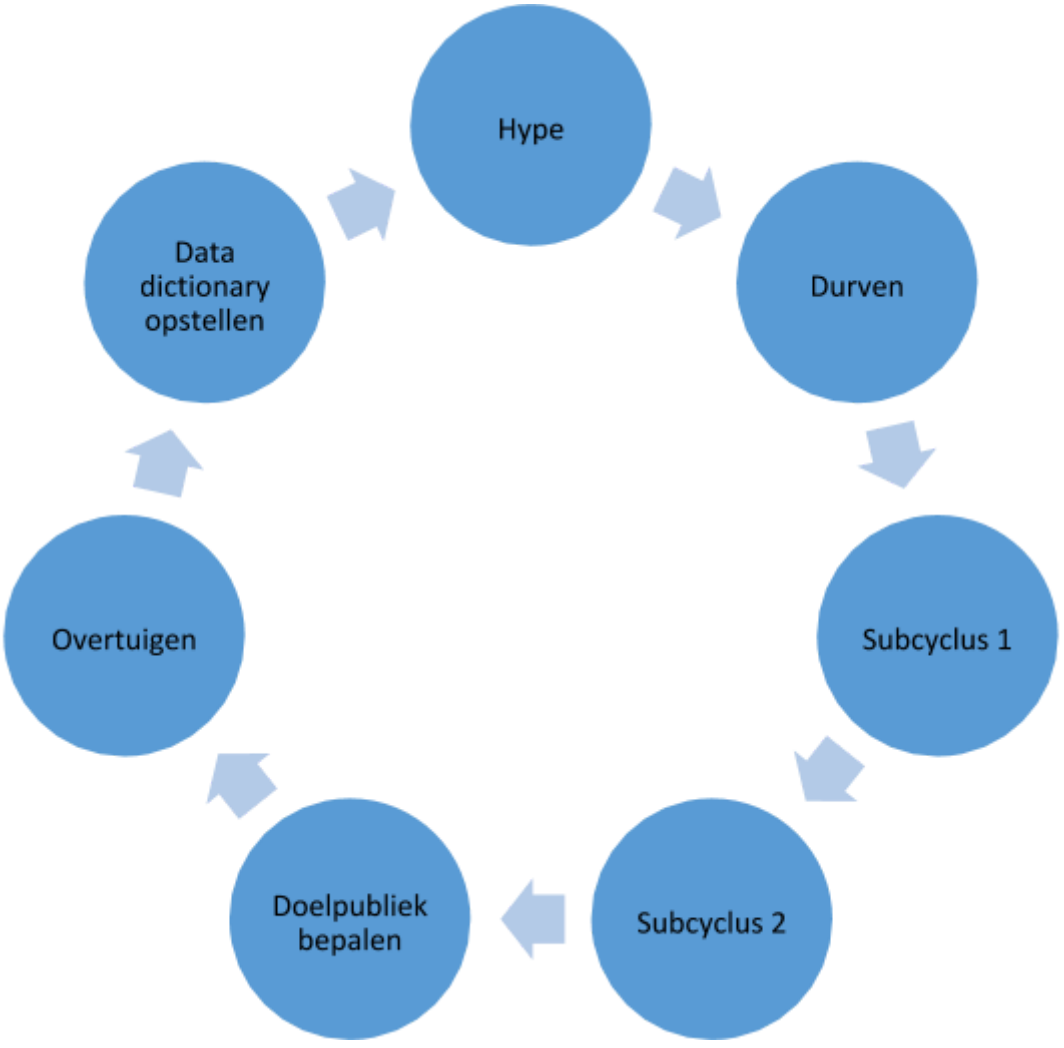
bepalen of één of meerdere van deze technologieën nuttig zouden zijn voor je organisatie.

De tweede stap is het aanwenden van menselijke input. Door te luisteren naar wat andere mensen hun gevoel van werking is, kan je het aantal te overwegen keuzes verder beperken.

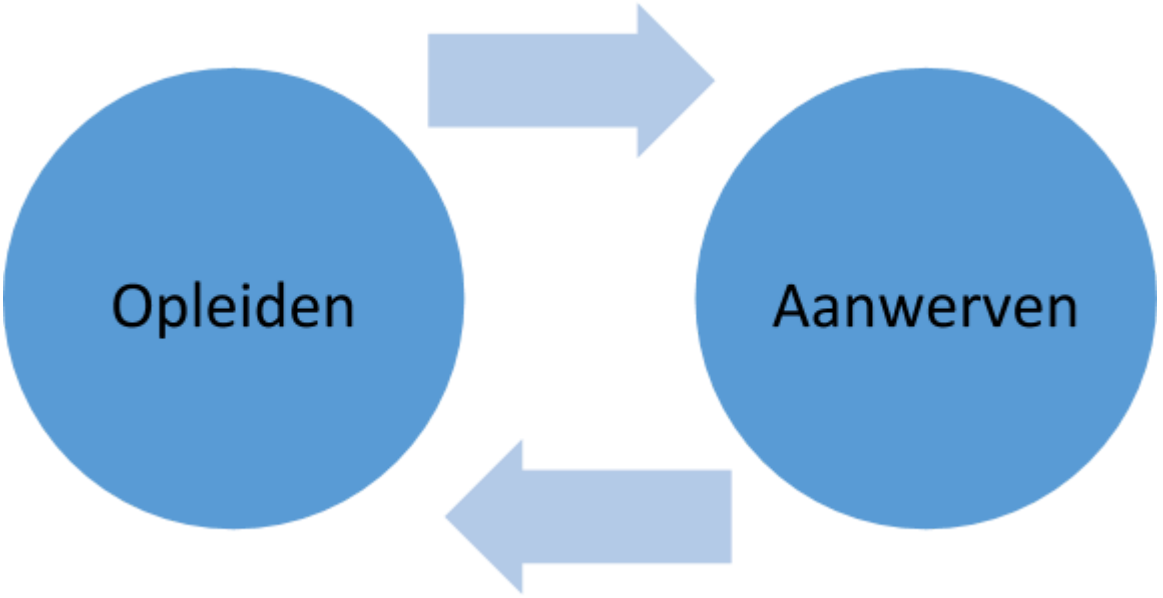
De derde en laatste stap is dan om in overweging hoe de te selecteren tool of technologie zou samenwerken en kan worden geïntegreerd met de reeds aanwezige technologieën (hadoop, SAS, …) en alle betrokken personen (business / IT).

5.4 Visuele voorstelling

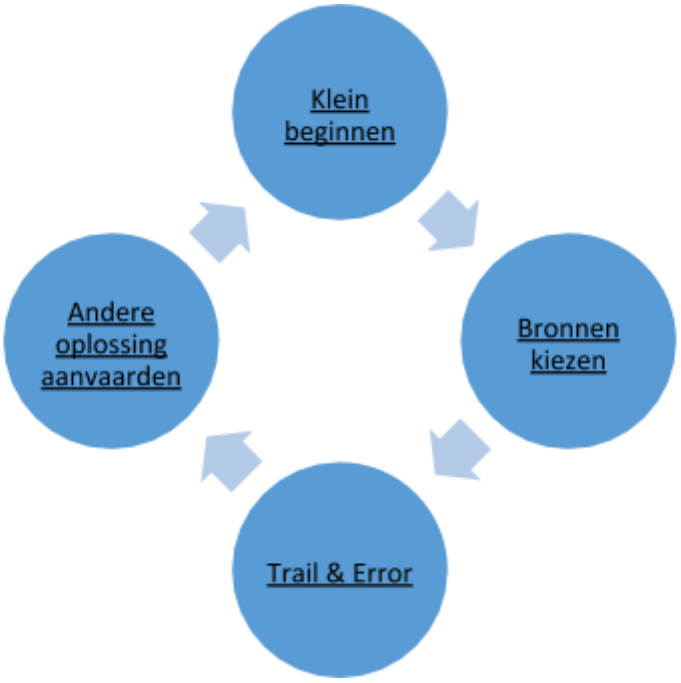
Levenscyclus



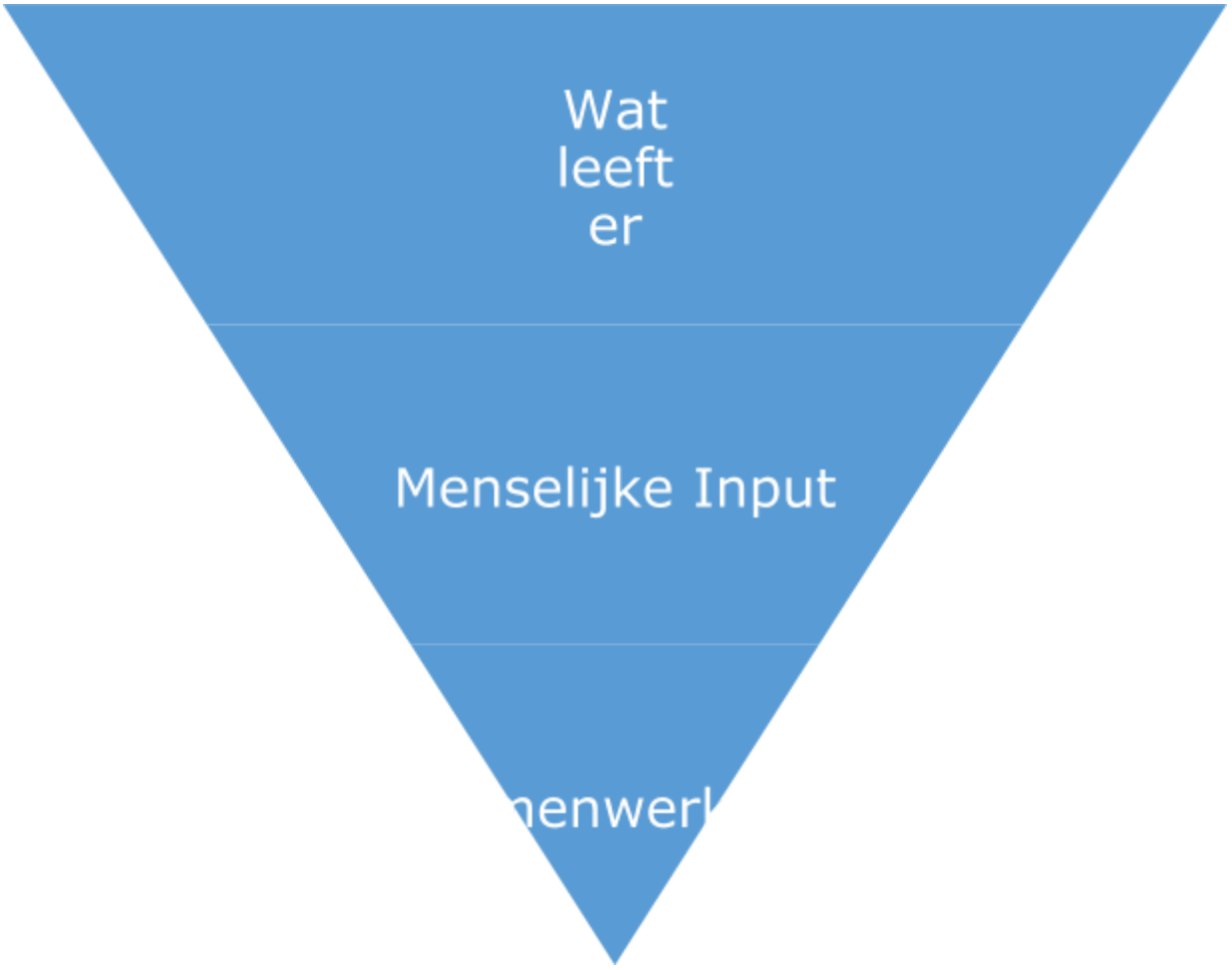
Sub cyclus 1



Sub cyclus 2



Trechtermodel



6. Proof of Concept

6.1 Omschrijving

Als proof of concept hebben wij gekozen om twitterstreams te gaan binnenhalen via Spark Streaming. We gaan hier alle tweets binnenhalen waar een van de top 10 merken van smartphones worden in vermeld. De top 10 smartphone producten zijn: Samsung, Apple, Huawei, Lenovo, Microsoft, Alcatel, LG, OneNote, Micromax, Xiaomi en ZTE. We gaan hier later via SAS Sentiment Analysis bepalen of een tweet positief of negatief is. Een van de best practices is klein beginnen. Daarom hebben we gekozen om deze use case uit te werken want dit is redelijk klein. We gaan ook onze Spark Streaming batches naar CSV-files laten wegschrijven. Deze CSV-bestanden worden dan nog eens opgekuist via RapidMiner. We gaan deze hierna wegschrijven naar een MySQL databank. We gebruiken een Oracle databank omdat onze data niet rechtstreeks en real-time verwerkt gaat worden. Hierdoor kunnen wij nog steeds een Oracle databank gebruiken. We gaan SAS laten gebruikmaken van onze MySQL databank, hier gaan we onze sentiment analyse en onze overige analyses op laten uitvoeren. We willen hiermee aantonen hoeveel er getweet wordt over een bepaald merk en meer bepaald wanneer. Voorbeeld als een bedrijf een nieuw product lanceert of aankondigt, wat de impact is op het tweetgedrag van hun fanbase en op het gedrag van de concurrentie.

6.1.1 Spark Streaming [23]

Spark streaming is een onderdeel van Apache Spark platform. Apache Spark is een open source processing engine dat gebouwd is rond snelheid, gebruiksvriendelijk. Spark streaming is een onderdeel van Apache Spark. Via Apache Spark Streaming halen we Twitterberichten op en gaan deze filteren, zodat we enkel onze top 10 smartphone merken door krijgen. Alle gerelateerde berichten worden opgeslagen in CSV-bestanden, waardoor we preprocessing en processing op een ander moment kunnen uitvoeren.

6.1.2 Databank aanmaken

We hebben geopteerd om een MySQL databank aan te maken. Dit is misschien niet volledig juist volgens het big data principe maar omdat we klein beginnen en geen real-time verwerking gaan uitvoeren leek dit ons een goed idee.

[23] Apache Spark, http://spark.apache.org/streaming, geraadpleegd op 23 januari 2017.

6.1.3 Data opkuisen

We gaan via RapidMiner onze CSV-files inlezen. Hierna gaan we de CSV-files bekijken kijken of er ontbrekende waarden, foutieve berichten, ... in de files bevinden. Tijdens het opkuisen van de data, gaan we ook onze data klaarzetten in onze MySQL databank.

6.1.4 SAS Contextual Analysis[24]

Tijdens deze stap gebruiken we gebruik maken van SAS Contextual Analysis tool. Hierdoor kunnen we in de Twitterberichten de tekst gaan analyseren of het bericht een positieve of negatieve boodschap bevat. We kunnen hier al een eerste inzicht gaan zoeken. Hier kunnen we ook onze data klaarzetten om later te gebruiken in SAS Visual Analytics.

6.1.5 SAS Visual Analytics[25]

Voor deze stap gaan we gebruik maken van SAS Visual Analytics gebruik maken. Hierdoor kunnen we op zeer eenvoudige manier aan onze MySQL databank koppelen en de data binnenhalen en analyses uitvoeren. Eén van de analyses dat we uitvoeren op onze data is over welk merk de meeste negatieve of meest positieve berichten worden gepost op Twitter.

6.2 Vergelijking met best practices

6.2.1 Hype

Omdat big data een hype geweest is en nu er gewoon is, hebben we na veel opzoekwerk besloten om hierrond een project op starten.

6.2.2 Durven

[24] SAS Institute NV/SA, http://www.sas.com/en_us/software/analytics/contextual-analysis.html, geraadpleegd op 26 januari 2017.
[25] SAS Institute NV/SA, http://www.sas.com/en_us/software/business-intelligence/visual-analytics.html, geraadpleegd op 23 januari 2017

Twee van onze teamleden moesten met dit project uit hun comfortzone treden omdat ze nog nooit iets met big data gedaan hadden. Ze durfden deze uitdaging aan.

6.2.3 Opleiden

De opleiding over big data en big data analytics hadden wij al deels verworven in de vakken "Big Data" en "Business Intelligence". Om dit project voor te bereiden hebben sommigen onder ons deze vakken in zelfstudie geleerd.

6.2.4 Aanwerven

Deze best practice was voor ons niet van toepassing aangezien wij niemand moesten of konden aanwerven.

6.2.5 Klein beginnen

Ons project hebben we klein genomen. We hebben namelijk gekozen voor producten die smartphones maken. We hebben ook gefilterd op smartphones indien ze meerdere producten maakten.

6.2.6 Bronnen kiezen

Wij hebben gekozen om twitter als bron te gebruiken. Hier zijn de meeste mensen aan het praten over de bedrijven.

6.2.7 Trial & Error

Dit is iets wat wij in de loop van het keuzevak "Big Data" al meegemaakt hebben. Hier hebben wij heel veel keer ons idee aangepast omdat het niet haalbaar leek.

6.2.8 Andere oplossingen aanvaarden

Dit is iets wat wij in de loop van het keuzevak “Big Data” al meegemaakt hebben. Toen we vast zaten dat onze leerkracht (Jan van Overveldt) ons een andere oplossing aanreikte die we aanvaard hebben.

6.2.9 Doelpubliek bepalen

Ons doelpubliek zal voornamelijk bedrijven zijn die hun concurrentie in kaart wil brengen. Bijvoorbeeld Samsung die weet wat hun klanten over hun producten denken, maar ook wat de impact bij de concurrenten zijn als ze iets nieuws releasen.

6.2.10 Overtuigen

Dit is een best practice die we niet kunnen aftoetsen, aangezien er op dit moment niemand is om te overtuigen.

6.2.11 Dictionary opmaken

Indien dit project volledig klaar en verkocht zou zijn, kunnen wij hier een data dictionary voor aanmaken. Dit voor het optimale gebruik.

6.2.12 Samenwerking

Door gebruik te maken van verschillende tools die met elkaar gaan praten, gaan we deze best practice proberen toe te passen.

6.2.13 Menselijke input

Het beste voorbeeld van onze menselijke input zijn de interviews die we gebruiken als input voor onze paper en onze proof of concept uit te werken.

6.2.14 Weten wat er leeft

Door meer en meer op te zoeken i.v.m. deze paper, weten we wat er leeft in de wereld van big data.

7. Besluit

Big data is niet meer weg te denken uit de huidige ICT-wereld. Het verzamelen en gebruikmaken van grote hoeveelheden data die in allerlei vormen pijlsnel over het internet vliegt, wordt een conditio sine qua non voor de concurrentiekracht van allerlei bedrijven.

Omgaan met deze data op een manier die je bedrijf sterker maakt, d.w.z. het vergaren van inzicht op basis van deze data d.m.v. big data analytics, is makkelijker gezegd dan gedaan. Om dit op een effectieve en efficiënte manier te kunnen doen, is het volgen van best practices noodzakelijk.

Het doel van deze paper was om een bevattelijk corpus van best practices voor big data analytics op te stellen, op basis van het advies van ervaren experts. Wij menen hierin geslaagd te zijn. Niet alleen hebben wij uit de combinatie van gangbare literatuur enerzijds en interviews met lokale experts anderzijds een lijst van 16 best practices gedistilleerd, tevens hebben wij deze best practices op een overzichtelijke manier gepresenteerd in de vorm van visuele modellen.

Een opvallend gegeven dat naar voren kwam tijdens het verwerken van de informatie die we vergaarden omtrent best practices, was de grote mate waarin de gevonden adviezen met elkaar overeenstemden. Dit zegt niet alleen iets over de validiteit van deze best practices, maar tevens over de maturiteit van het domein. Ondanks het relatief recent ontstaan van big data analytics, lijken de experts het eens te zijn over welke richtlijnen gevolgd dienen te worden.

Een mogelijke piste voor vervolgonderzoek zou zijn om de door ons geconstrueerde modellen voor te leggen aan experts ter controle van hun validiteit, en om de samenhang tussen de verschillende best practices (zowel op louter theoretisch vlak als in de context van een geheel van bedrijfsprocessen) verder te onderzoeken.

8. Bronnen

8.1 Afbeeldingen

Figuur 1: Big Data, http://www.allezorgvergoedingen.nl/images/big_data.jpg, geraadpleegd op 26 januari 2017.

Figuur 2: Apache Hadoop logo, http://hadoop.apache.org/images/hadoop-logo.jpg, geraadpleegd op 26 januari 2017.

Figuur 3: Descriptive tot prescriptive, http://www.rosebt.com/uploads/8/1/8/1/8181762/6779091_orig.jpg, geraadpleegd op 26 januari 2017.

Figuur 4: Logo SAS, http://www.sas.com/en_be/legal/logos.html, geraadpleegd op 26 januari 2017.

Figuur 5: Logo Big Industries, http://www.bigindustries.be/, geraadpleegd op 26 januari 2016.

Figuur 6: Logo Infofarm, http://www.infofarm.be/, geraadpleegd op 26 januari 2017.

Figuur 7: Logo I4BI, http://i4bi.be/, geraadpleegd op 26 januari 2017.

Figuur 8: Logo Deloitte, https://www2.deloitte.com/content/dam/Deloitte/il/Images/promo_images/deloitte_logo_black_1x1.jpg, geraadpleegd op 26 januari 2016.

Figuur 9: Logo Spark Streaming, https://www.mapr.com/sites/default/files/otherpageimages/spark-streaming.png, geraadpleegd op 26 januari 2017.

8.2 Cursussen

VAN OVERVELDT J., 1_Introductie, Antwerpen, Karel De Grote-Hogeschool, l.d., 28.

VAN OVERVELDT J., 3_DataGovernance, Antwerpen, Karel De Grote-Hogeschool, l.d., 20.

VAN OVERVELDT J., 5_DataBeschikbaarMaken, Antwerpen, Karel De Grote-Hogeschool, l.d., 3.

VAN OVERVELDT J., Big Data Introductie, Antwerpen, Karel De Grote-Hogeschool, l.d., 2.

8.3 Interviews

CASSIERS H., *Mondelinge mededeling* (interview), Kontich, 18 januari 2016, 10.00 uur.

RENDERS E., *Mondelinge mededeling* (bijlage in mail), Kontich, 15 januari 2016, 13.30 uur.

SEVEREYENS A., *Mondelinge mededeling* (interview), Tervuren, 18 januari 2016, 13.30 uur.

VALLAEY M., *Mondelinge mededeling* (interview), Kontich, 17 januari 2016, 13.30 uur.

VAN DEN BROECK M., *Mondelinge mededeling* (interview), Kontich, 17 januari 2016, 13.30 uur.

VAN DYCK T., *Mondelinge mededeling* (interview), Tervuren, 18 januari 2016, 13.30 uur.

VERMEERSCH B., *Mondelinge mededeling* (interview), Kontich, 17 januari 2016, 10.00 uur.

VERSAILLES E., *Mondelinge mededeling* (interview), Tervuren, 18 januari 2016, 13.30 uur.

8.4 Websites

APACHE HADOOP, http://hadoop.apache.org/, geraadpleegd op 23 januari 2017.

APACHE SPARK, http://spark.apache.org/streaming, geraadpleegd op 23 januari 2017.

BIGDATAWEEK, http://blog.bigdataweek.com/2016/02/23/keys-big-data-start-small-think-big-grow-fast/, geraadpleegd op 26 januari 2017.

COMPUTERWEEKLY, http://www.computerweekly.com/podcast/Big-data-storage-Defining-big-data-and-the-type-of-storage-it-needs, geraadpleegd op 23 januari 2017.

DATA GOVERNANCE INSTITUTE, http://www.datagovernance.com/the-basic-information/, geraadpleegd op 23 januari 2017.

INFORMATIONWEEK, http://www.informationweek.com/big-data/big-data-analytics/structuring-your-data-team-9-best-practices/d/d-id/1323601?image_number=6, geraadpleegd op 26 januari 2017.

JAMESDIXON'S BLOG,

https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/, geraadpleegd op 23 januari 2017;

MICROSOFT, https://msdn.microsoft.com/en-us/library/bb190163.aspx, geraadpleegd op 23 januari 2017.

QUALITY BUSINESS SUPPORT BLOG, https://qualitybs.wordpress.com/2012/01/29/vastleggen-van-best-practices/, geraadpleegd op 24 januari 2017.

SAS INSTITUTE NV/SA, http://www.sas.com/en_us/software/analytics/contextual-analysis.html, geraadpleegd op 26 januari 2017.

SAS INSTITUTE NV/SA, http://www.sas.com/en_us/insights/analytics/big-data-analytics.html, geraadpleegd op 23 januari 2017.

SAS INSTITUTE NV/SA, http://www.sas.com/en_us/insights/big-data/what-is-big-data.html, geraadpleegd op 23 januari 2017.

SAS INSTITUTE NV/SA, http://www.sas.com/en_us/software/business-intelligence/visual-analytics.html, geraadpleegd op 23 januari 2017.

TECHOPEDIA, https://www.techopedia.com/definition/28966/data-sandbox-big-data, geraadpleegd op 26 januari 2017.

8.5 White papers

M. BARBERO, e.a., *Big data analytics for policy making*, European Union, 2016, 122 p.

X, *An Enterprise Architect's Guide to Big Data*, Oracle Enterprise Architecture White Paper, maart 2016, 45 p.

8.6 Magazineartikels

X, "5 best practices for big data analytics", *Network World Asia*, nov/dec 2015, p. 6.

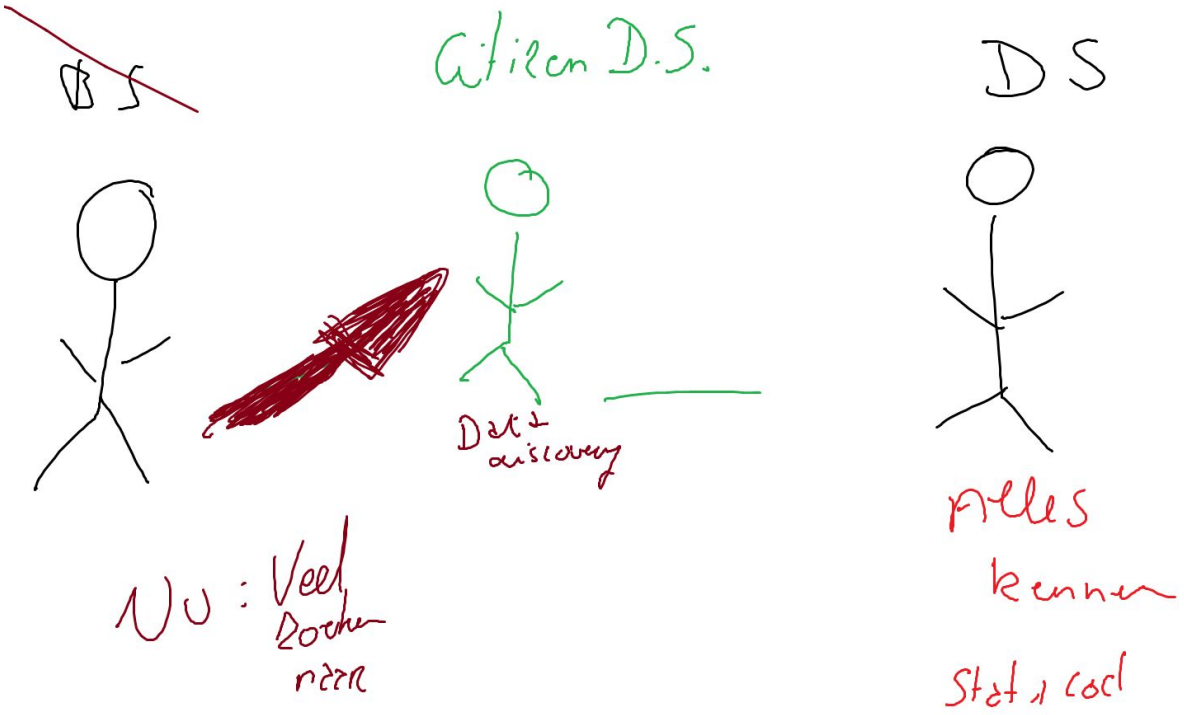
9. Bijlagen

9.1 Interviews

9.1.1 Interview I4BI

Algemeen

1. Hoelang bent u al bezig met Big Data of Business Analytics?
   13 jaar : DWH / BA
   2 jaar: Big data (theoretisch weinig, meer opvolging van wat leeft er en wanneer doen + vorig jaar stage)
   2 maand (sinds half november)
2. Wat hebt u gestudeerd of hoe bent u bij Analytics uitgekomen?
   Doctoraat biologie + hackaton ; verteld over doctoraat , sloot aan bij hackaton en zo begonnen.
   Industrieel ingenieur electronika optie ICT (KDG) – gestart bij consulting firma en opleiding in transactioneel systeem gekregen en volgend project was BI … 10 j
   Sinds vorig jaar XploData.
3. Zijn er genoeg mensen in België die zich bezig houden met Analytics?
   Te weinig, eerder krap op de markt.
   Redelijk moeilijk om juiste profielen te vinden waarbij puur rapporteringsluik minder vraag is.
   Advanced analytics is de vraag er wel
   DWH zeker krappe markt.
   2/3 mensen rapportering vs 10 mensen data verwerkingen
   Data scientisten zijn nog krap (business kant en IT kant + communicatie)
   Er zijn mensen die volledige scientist zijn.
4. Is BA / DS echt iets voor business mensen of eerder voor meer IT mensen?
   Mix van beide. Je hebt mensen nodig met business achtergrond, de hacking skills / IT skills
   Meer en meer softwareproviders komen af met tools voor Citizen Data Scientists (zijn mensen die niet altijd wiskundige achtergrond hebben ; dus Data scientists nog steeds nodig)
   Data scientists kunnen citizen data scientist ondersteunen.
   Belangrijk blijft input van bedrijfsleven.



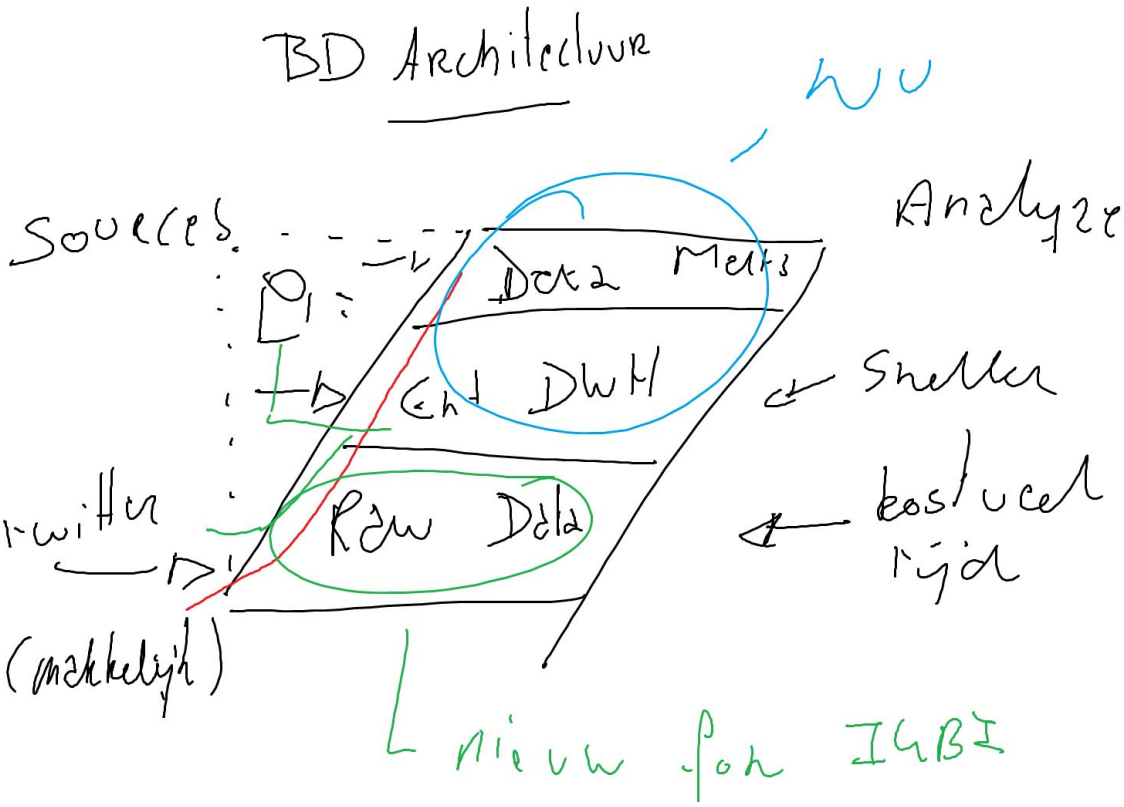
5. Als je een data scientist zou moeten beschrijven in 3 woorden, welke zouden deze dan zijn?
   Statistiek: kennis van data en wat er mee doen
   Business: hoe het bedrijf werkt / bezig zijn. Met een frisse blik naar de zaken kunnen zien.
   It/Hacking: Computer skills
   (Communicatie).

6. Is Big Data en Data Analytics een hype of eerder nog steeds iets veel belovend?
   Het was zeker een Hype. Er is al veel geld verloren gegaan aan platformen die opnieuw gebouwd moeten worden.
   In België komt het geleidelijk aan.
   Overheden heeft nu heel veel vraag (concreet bezig).
Farmacie bedrijven en telecom bedrijven zien hier al aan bezig.
   VDAB is matuur bezig op Big Data en zitten volledig in google / google drive en werken Agile.

7. In welke onderwerpen/opdrachtsoort hebt u het meeste ervaring?
   Jansens farmaceutica: DWH
   Oracle BI rapportering tool
   Oracle Visual Analyser/Tableau

Big Data

8. Met welk platform hebt u het meeste ervaring mee of raadt u het meest aan (Big Data)?
   Terradata: Aster-platform (Big Data van Terradata – query's loslaten op data)
   Aster kan naar alle verschillende db'en connecteren, en aster gaat analyses zelf uitvoeren.
   Cloudera
   Oracle Big Data Appliance
   HortonWorks
   Aster DB heeft goede relatie met R



9. Wat is belangrijk bij het verzamelen van data?
   Trail and error (binnen halen en valideren) / Fail fast and retry (discovery track)
   Zoveel mogelijk data op dataplatform te plaatsen.
   Iets wat nu niet waardevol is kan wel waardevol worden
   Bron identificeren waar interesse voor is ; gebruik geen subset maar neem alles mee (volledige scope)

10. Wat zijn de belangrijkste use cases (Big Data)?
    Business opschuiven naar Data Scientist verhaal
    (VDAB heeft werkwinkels voor als je werk zoekt, website/platform is groot aan het worden en wordt beter en minder en minder mensen komen naar de winkels – kijken naar logs om valideren welke mensen er komen)

    Vlaamse overheid VO : Impact bekijken van beslissingen van minister

    J&J: in kaart brengen waar de hotspots zijn van uitbraken van virussen. Om hulpstelling beter te kunnen concentreren. Twittersstreams analyseren, verschuivingen nakijken van dieren/mensen, weerscondities, … In groot model een globaal en regionale voorspellingen van uitbraken.

    Staalproducent: machinepark met pletten en walsen vermoeden dat ze veel energie verspillen door slecht afgestelde machines. (Data discovery op uitvoeren) Is er genoeg data om analyses te doen. Zien waar logging moet uitgebreidt worden.

11. Wat zijn de belangrijkste Big Data Technologieën?
    Hadoop, Aster DB, cloudera
    ETL tools: die raw data kunnen volgen (INFA / ODI) – connectoren naar alle platformen – creëren big data technologieën, om stabiel omgeving te creëren voor Big data.
    Als ETL niet kan – Data Scientist

12. Hoe moet een organisatie zich aanpassen voor Big Data?
    Architect (of team) moet heel plaatje beheren (zie vorige afbeelding)
    Meeste kijken enkel naar ruwe Data maar ze moeten ook kijken naar de andere lagen (ENT DWH en DATA MARTS)
    Architect moet rekening houden met alles dat al bestaat.
    BI mensen kijken enkel naar Data Marts. (krijgen wel tools die over alle lagen gaan)
    DWH mensen: ETL die kijken enkel naar ENT DWH laag; moeten connectie met Raw Data bijleren. (leren querieën) MAAR moeten ook basiskennis opbouwen (basis file transformatie) Bijscholen in Java / dotNet.
    BD developer: zie Infofarm
    Business kant: blijven analyses doen (standaard BI is er al) , gaan meer naar een Citizens Data Scientist gaan.
    Data scientist met toegevoegde waarde en samenwerking.
    Big data developers bij aannemen.
    Gap tussen data scientist en citizen data scientist is nog te groot.

13. Welke invloed heeft Big Data op een bedrijf?
    Business value uithalen als je het goed doet (voorbeeld van staalbedrijf is goed voorbeeld)
    grootste fout : veel te veel tijd in platformen gestoken zonder te weten wat je nodig hebt.
    (log file niet juist: komt er na 2 jaar achter ➔ Verkeerd)
    Wat doe je, en waarom doe je het. Voor je begint de data toch al nakijken.

Big Data Analytics

14. Met welk platform hebt u het meeste ervaring mee of raadt u het meest aan (Analytics)?
    Oracle big data discovery tool, Terradata, Oracle BI (dashboarding), R, Visual Analyser

15. Wat zijn de belangrijkste zaken om op te letten om een analytics platform op te stellen?
    De juiste keuzes maken (juiste tools kiezen)
    De juiste tool kiezen voor de juiste use case.
    Oracle heeft BI publisher (pixel fine werken) voor printouts en doorsturen naar bedrijven ➔ niet goed voor standaard programmeringstools.
    Juiste keuze qua infrastructuur.

Doelpubliek bepalen en hiervoor keuzes maken.

16. Wat zijn de belangrijkste use cases (Analytics)?
    zie vraag 9

17. Welke invloed heeft Analytics op een bedrijf?
    Als je het goed doet kan je bepaalde zaken berekenen. Dagdagelijkse opvolging. Bonussen berekenen.
    Maar nu berekeningen ; waarom haalt iemand zijn targets niet.
    op termijn: Wat is de invloed van machine tweeking en opvolgen.
    Klopt model als je machines vervangt, energieverbruik
    daarna: Predictive analyses doen.
    automatisch beslissingen laten nemen door de modellen.
    Zijn sales rep goed bezig, of zijn ze met te veel. (werkwinkels van VDAB – website voldoen en sluiten)
    Kan grote invloed hebben als je het goed doet.

Slot

18. Hoe plan je een roadmap? Wat zijn de Best Practices en zaken waar we op moeten letten?
    Afbeelding van de parallellogram en hoe aanpakken.
    Afhankelijk van de klant (wat willen ze doen nu, toekomst …) Op die basis gaan bepalen welke laag je nodig hebt.
    Zeer kritisch staan tegenover wat de noden zijn en hoe gaan we die kostefficiënt invullen.
    Kunnen alles ineens doen maar helft niet nodig.
    Fit op purpose stuff maken op basis van input van de klant.
    In bepaalde gevallen kan je niet rond het platform maar wil niet zeggen dat je daar dan analyses op moet doen
    luisteren naar wat de klant nodig heeft.
    one size fits all.
19. Zijn er nog zaken die wij niet hebben aangehaald hebben die belangrijk zijn om te vermelden.
    Grote wiel is besproken, alles is wat aangehaald.
    Big data mag je niet als 1 blokje zien maar als geheel.
20. Zijn er specifieke bronnen over Big Data Analytics die u ons aanraadt die ons nog zouden kunnen helpen?
    Data science boek (2013) : geen hype maar werkelijkheid
    afbeelding van platformen: whitepaper over
    terradata architectuur slide.

9.1.2 Interview Big Industries

Algemeen

1. Hoelang bent u al bezig met Big Data of Business Analytics?
   4 jaar.
2. Wat hebt u gestudeerd of hoe bent u bij Analytics uitgekomen?
   Master in Toegepaste Economische Wetenschappen
   23 jaar werkervaring , altijd in de IT (Vendors)
   4 jaar geleden bij Cronos begonnen, vanuit ondernemerschap
   toen was Big Data een hype maar niemand wist hoe er aan te beginnen.
   Cronos wil altijd direct meestappen in nieuwe tech
   Hierdoor is hij begonnen met op Cronos niveau big data te starten
   Kijken hoe je big data project analyseert. Door trainingen en gericht gaan werken.
   3 jaar geleden eerste klant Telenet.
   Levensvatbaar dus daarmee is Big Industries begonnen.
3. Zijn er genoeg mensen in België die zich bezig houden met Analytics?
   Meer en Meer , 4 jaar geleden was er niemand maar nu veel meer mensen.
   3 jaar geleden (Robert mede oprichter)
   Robert heeft in 2009 eerste hadoop cluster opgericht.
   Het heeft een trage start gemaakt.
   Er is langs ene kant te kort aan Big Data en aan Data Scientist maar niet meer als andere (zoals dotNet)
   Gap kan opgelost worden door mensen die het willen doen die een stap willen zetten van dev naar big data. Gap wordt ingevuld.
4. Is BA / DS echt iets voor business mensen of eerder voor meer IT mensen?
   Kruispunt van de 2.
   Bij het begin bij de hype en mystiek van data science een project voor aquafin: research budget , ze gaven een dataset en moesten een datascience project oplossen en ze moesten met data komen dat ze nog niet wisten. De ingenieurs wisten al lang waar dat ze mee kwamen.
   Het is technisch (achtergrond statistiek, wiskunde, computerwetenschappen) en bedrijfskennis.
   Het is een kruispunt tussen de 2.
5. Als je een data scientist zou moeten beschrijven in 3 woorden, welke zouden deze dan zijn?
   Datamining
   Statistiek

Business kennis

6. Is Big Data en Data Analytics een hype of eerder nog steeds iets veel belovend?
   Big Data is over zijn hype heen, maar is nog wel veel belovend. In het hedendaags leven is het een dood normale zaak. Als je kijkt naar de grote bedrijven gebruiken allemaal al big data (zoals. mediabedrijven, farmacie, public sector, banken, ...).
   Vlaamse overheid is van datacenter naar Amazon gegaan. (concept van Data Lake)
   Alle start up community gebruikt big data
   Wat zit er nog niet op : KMO (wat willen ze er mee doen / maturiteit).
7. In welke onderwerpen/opdrachtsoort hebt u het meeste ervaring?
   Data lake, het vervangen of van scratch opzetten van DWH / Data Lakes.
   Andere data bronnen dan Databronnen (geen sensor , geen klik)
   Principe blijft hetzelfde (pipeliness om alle data in db's te krijgen) maar de data verschilt.

Big Data

8. Met welk platform hebt u het meeste ervaring mee of raadt u het meest aan (Big Data)?
   On premise: Cloudera
   Cloud: Amazon
   zijn de 2 marktleiders
9. Wat is belangrijk bij het verzamelen van data?
   Bruikbare data, relevante data
   Het heeft te maken met de use case, wat is u use case en welke data heb je hier voor nodig.
10. Wat zijn de belangrijkste use cases (Big Data)?
    Operational predictive/prescriptive analytics: last van kwaliteit kabelmodems, 6 maand logdata bij elkaar halen om zo te voorspellen wanneer modems kapot zouden zijn. (met mysql is na week db gecrasht).
    Security: Intrusion Detection system die logs bijhouden (veel data) die meestal wordt weggesmeten. In de logfiles kan je achterhalen of er een lek geweest is dat je niet gemerkt hebt.
    Marketing Customer 360: (telenet/mediahuis): Steeviee/vtm platform om alle logfiles binnen trekken. Door account weten we exact wie je bent. Alles wordt gelogd. Vtm kan zo bepalen welk seizoen van welke serie te kopen. Target advertising, verder kijken.
    Volgende klant: kranten
    Mobile Vikings
    ➔ *Wat kijkt de mens, wat leest de mens, waar is de mens*
11. Wat zijn de belangrijkste Big Data Technologieën?
    Hadoop: 95 % van de projecten
    Spark
    NoSql DB (Cassandra) en equivalente voor AWS
    Graph DB
    Amazon
    Phyton
    Data science: R / Python / Spark

12. Hoe moet een organisatie zich aanpassen voor Big Data?
    Opleiding in de technologieën.
    Openstaan voor innovatie. De nood voelen dat er iets moet veranderen om te groeien.
    De grootste concurrent van TUI is niet neckermann maar booking.com
    De grootste concurrent van Persgroep/Mediahuis is facebook
    Nieuwe bedrijven die de markt veroveren die gebruik maken van Big Data.
    Klassieke bedrijven moeten antwoord hebben : INNOVATE OR DIE
    De rest volgt wel.
13. Welke invloed heeft Big Data op een bedrijf?
    De mogelijkheid om te innoveren.
    Kostenbesparing: uitgerekend als ze voor klassieke Oracle db te doen vs hadoop
    alles blijft hetzelfde maar licenties: Oracle : voor big data en na big data was het verschil 10 keer zo weinig (van M bedrag naar K bedrag)
    Moest hadoop niet bestaan zou het project van vtm nooit doorgegaan zijn.

Big Data Analytics

14. Met welk platform hebt u het meeste ervaring mee of raadt u het meest aan (Analytics)?
    De klassieke datamining: Sas & SPSS
    nieuw: R en Python
    Data science project in context van Big Data 90% gevallen R & Python
15. Wat zijn de belangrijkste zaken om op te letten om een analytics platform op te stellen?
    Wat moet er allemaal inzitten: interfaces om data er in te krijgen
    Data management & Data governance project: data lineage, master data management, security (securityregels dat niet iedereen naar de data kan) , de juiste tools om data te preparen, visualisatie

16. Wat zijn de belangrijkste use cases (Analytics)?
    Fraudedetectie: voor public sector klant (SMALS) stellen ze nu big data analytics platform en ze willen alle verschillende db's bepaalde patronen gaan zoeken
    europese commissie : weging van containers (btw caroussels detecteren)
    Life Science research: Genen onderzoek
    Costumer 360: VTM
    predictive analyses.
17. Welke invloed heeft Analytics op een bedrijf?
    Betere inzichten; als je dit verstandig gebruikt; Operationele processen verbeteren.
    Telenet; oude manier werken: kabelmodem kapot – bellen – wachten en geraakt geirriteerd – kan tot 3 dagen duren da ze kunnen langskomen
    predictive: kans dat modem kapot gaan -> brief sturen om modem aan te passen.
    Operationele teams kunnen werk plannen, tevreden klanten en efficiënter werken.

Slot

18. Hoe plan je een roadmap? Wat zijn de Best Practices en zaken waar we op moeten letten?
    Beginnen met een goede use case (wat wil je bereiken, wat wil je doen)
    kijken naar data bronnen (betrouwbaar/ goede kwaliteit)
    Technologiekeuze (juiste tool voor juiste job)
    Overtuigen: Goede visualisatie
19. Zijn er nog zaken die wij niet hebben aangehaald hebben die belangrijk zijn om te vermelden.
    Allemaal nieuwe technologie die heel veel veranderd; je moet gewoon durven en uit je comfortzone gaan, experimenteren en durven op je bek te gaan.
    Durf nieuwe paden te zoeken. (forums in gaten houden, weten wat er leeft in de big data wereld)
20. Zijn er specifieke bronnen over Big Data Analytics die u ons aanraadt die ons nog zouden kunnen helpen?
    Forums over hadoop, etc…
    hadoop essentials (hoe ontstaan, alle componenten uitleggen, naslagwerk)
    O'reilly : definitive guide (bijbel van Big Data)
    training.cloudera.com : cloudera essentials for apache hadoop (204 slides , maar goed)

9.1.3 Interview Infofarm

Algemeen

1. Hoelang bent u al bezig met Big Data of Business Analytics?
   Persoonlijk / Infofarm: 2014 begonnen. Ontstaan vanuit technisch idee (java).
   Cronos: 300 tal bedrijven (4-5 K werknemers). Bedrijven focussen op eigen ding, maar hebben hun eigen core.
   Verschillende subgroepen (Xplore,…).

Xplore groep: maatoplossingen softwareontwikkelingen (Big data / Big data Science)
Begonnen van uit technisch standpunt maar wilden wel iets mee Big Data doen.
Vanuit combinatie big data en science is infofarm ontstaan.
2015: zelfstandig onder Xplore/Cronos groep.
3 tal jaar specifiek rond big data en big data science

2. Wat hebt u gestudeerd of hoe bent u bij Analytics uitgekomen?
   KdG – Toegepaste Informatica (Applicatieontwikkeling)

2 profielen in infofarm: technisch (ontwikkelaars): Toegepaste Informatica (kennis van programmeren – java, scala, kennis van applicatieontwikkeling, spring, cloud, angular, big data goed meegenomen)
Achtergrond ontwikkeling belangrijk
Data scientist: Kwantitatief veld, wiskunde, biomedisch, statistisch (Universiteit)
2 Uiterste. Maar daar tussen kan ook in aanmerking komen
soft skills (praten) heel belangrijk.
Niet weten hoe , maar waarom en waar implementeren.
50 – 50 Science / Devs
Geloven meer in teams van mensen dan iemand die alles zelf kan.

3. Zijn er genoeg mensen in België die zich bezig houden met Analytics?
   Heel veel mensen die zich bezig houden, maar is een heel breed spectrum.
   BI bestaat al jaren en is niet nieuw (veel mensen mee bezig)
   ML bestaat ook al jaren.
   nieuw: Big data tooling (cloud , hadoop, …) -> nieuwe dynamiek en meer vraag, bereikbaarder.
   De kennis rond Data Science / Big data science is nog niet hard verspreid (nog veel nodig).
   Veel mensen mee bezig , en heel veel komen er bij.
   Infofarm: niet steunen op externe tools, maar maatoplossingen. Niet afhankelijk van drag & drop tool.
4. Is BA / DS echt iets voor business mensen of eerder voor meer IT mensen?
   Om data science te doen of gegevens gebruiken moet je heel veel kennis hebben van het bedrijf (weten wat leeft, waar zitten problemen, ….) (vb. Aquafin: water dat loopt door rioleringen analyseren voor storting te kunnen controleren – voorspellen waar het begin te regenen, hoe gaat het water door de riool vloeien en wanneer is er een probleem en water andere richtingen op sturen)
   Business kennis is nodig. Business kennis is groot maar modellen is minder. Technische kennis nog voor producten te maken. In teams oplossen.
   R wordt bij infofarm gebruikt. Hangt van factoren af of ombouwen tot Spark of gewoon interface boven R zetten.
   Vertalen naar bedrijfstaal.
   belangrijk is communicatie, wat leeft er in een bedrijf, hoe breng je resultaten over.
   vb. Esser: van waar naar waar rijdt een camionne en waar rijden er veel leeg.
   Validatie en interpretatie gegeven heel hard business voor nodig (welke contracten hebben ze) heel veel tools komen online, om business mensen het te laten doen. Tools beter en beter, kunnen meer met grotere data omweg. Achterliggend weten ze niet hoe het werkt.

analytische gevaren: IBM Watson – mooi en gebruik van dataset en inzichten halen. Rondklikken en analyses uithalen (waarom verlaten mensen het bedrijf, …) kan nuttig zijn maar gevaarlijk omdat je niet weet hoe de tool tot de beslissingen komen. (vertegenwoordiger die er meer in komt dan andere, …)
Heel veel gevaren verbonden omdat je niet weet hoe ze er bij komen (daarvoor data scientist nodig) .

5. Als je een data scientist zou moeten beschrijven in 3 woorden, welke zouden deze dan zijn?
- Analytische kennis : hoe omgaan met datasets, welke modellen kunnen hier met om, hoe data omzetten
- Communiceren: Er zijn mensen die prachtige dingen kunnen doen met hun data maar niet kunnen uitleggen wat ze ermee kunnen doen. Je moet kunnen uitleggen wat een model doet.
- Kennis van programmeren: nuttig om breder inzetbaar te zijn (prototype maken / interface maken) (hacken om data te verzamelen, data verzamelen)
6. Is Big Data en Data Analytics een hype of eerder nog steeds iets veel belovend?
Het is een hype , alle bedrijven moeten meedoen of je bent niet mee. Nadeel dat er veel bedrijven opspringen dat er niets van kunnen. Het is matuurder dan vroeger (door meer mensen, meer integratie , meer software, stabielere platformen, Hadoop van mapreduce naar top level tools (hive, pig , tez, impala, storm) nadeel : allemaal eigen interface. Naar Spark: 1 interface die alles doet → Werkt gemakkelijker)
Heel veel bedrijven moeten er nog mee beginnen en het is nog steeds veelbelovend.
7. In welke onderwerpen/opdrachtsoort hebt u het meeste ervaring?

Heel breedt, meest gebeurd: Uitbreiding van BI / DWH omgeving, Integratie met bestaande programma's , importen naar big data technologieën waarop ze dan verder gaan werken. De sector is redelijk breed aan opdrachten.
Aquafin: puur analyse
trein bedrijf: analyse en uitwerken
Recommendation engine voor webshop, is klant een koper of niet.
neuraal netwerk maken voor image recognition
Transport bedrijf
Vuilnis ophaling
HR …
Met veel technieken kan je meerdere dingens doen. (voorspellen of klant vertrekt of niet: Turn prediction)Turn prediction: klant weg, locomotief kapot, ….

Big Data

8. Met welk platform hebt u het meeste ervaring mee of raadt u het meest aan (Big Data)?
Hadoop / Spark: standaard voor big data te verwerken, open source, zeker voor cluster computing DIY: Hadoop installeren from source, heel moeilijk
Commerciele vendors: Cloudera, Hortonworks (prepackaged hadoop) met toplevel tools, Oracle Big Data appliance (cluster kopen) (stekker insteken en het staat op)

Cloud: AWS, IBM, Azure

9. Wat is belangrijk bij het verzamelen van data?
Niks wegsmijten, alles bijhouden want in traditionele BI db met alles in schema wordt veel weggegooid. Met Big data alles bijhouden, opslagen hoe ge hebt binnen krijgt en geen preprocessing doen. Data dump aanleggen. Net geen data lake aanleggen want hier staat nog is een interface op. hierna kan je alle ruwe data omvormen / transformeren.
10. Wat zijn de belangrijkste use cases (Big Data)?
Kan redelijk breed gaan; Data processing + Rapportering , data verwerking als hub tussen 2 systemen (etl), voorspellingen voeren, Operationele zaken ondersteunen.
Vereiste: Veel data hebben of weten dat het veel data kan worden (scaling)
11. Wat zijn de belangrijkste Big Data Technologieën?
Hadoop eco systeem is het belangrijkste.

12. Hoe moet een organisatie zich aanpassen voor Big Data?
    Infofarm bellen :p Development is nodig , er zijn heel veel tools op de markt en er komen er meer en meer aan. Dit wil niet zeggen dat deze vele tools op de markt gelijk staat aan goede tools. Mensen die drag & drop kennen zijn beperkt tot die tool. Als er iets misgaat weet je niet hoe het komt en hoe op te lossen. Technische kennis is nodig. (hadoop omvormen, interface schrijven, …)
13. Welke invloed heeft Big Data op een bedrijf?
    Dingen die moeilijk waren vroeger, die nu ineens heel makkelijk worden. Data werd weggegooid of gestored, en opzoeken duurde lang. Nu simpele querie afvuren en makkelijk vinden. Niet verwerkbare gegevens kunnen nu wel verwerkt worden. Niet zelf meer gaan zoeken hoe je gegevens moet verspreiden over verschillende systemen

Big Data Analytics

14. Met welk platform hebt u het meeste ervaring mee of raadt u het meest aan (Analytics)?
    Eerder minder ervaring / gebruiken minder tools hiervoor. Meest gebruikte tool is R.
    Dit wordt op maat van de klant gemaakt. Gebruiken geen tools om analytics op te doen.
    ElasticSearch met Kibana voor snel te querieën en veel gegevens in te steken om data eens te bekijken / aan business te geven. Om zo te zien of er informatie in te winnen is.
    Vooral eigen ontwikkeling, en de programmeer technische mensen vormen het om naar tools. Oppassen voor gelimiteerde gegevens. Oppassen voor tools waar je niet weet hoe het werkt achterliggend.
15. Wat zijn de belangrijkste zaken om op te letten om een analytics platform op te stellen?
    Technisch gezien: Eender wat je gebruikt, weten wat je rapporteert, weet welke analyses je maakt, weet wat je data wil zeggen, weten wat er achter zit.
    Hierdoor werken zij alles zelf uit dat ze zelf weten dat alles juist zit.
16. Wat zijn de belangrijkste use cases (Analytics)?
    zie vraag 10.
17. Welke invloed heeft Analytics op een bedrijf?
    *Heel veel **buikgevoel bevestigen of ontkrachten**. Momenteel wordt er veel van buikgevoel beslissingen genomen die veel juist zijn maar soms helemaal niet meer. Kennis bij 1 iemand in het bedrijf kan opgelost worden door puur cijfermatig naar gegevens te gaan kijken.*
    Essers: De kennis dat vrachtwagens leeg reden was 1 iemand die da wist. Maar niemand anders.
    Fashion Retailer: beste locatie bepalen + invloed online sales als er een winkel in een regio zit. (korte termijn boost)

Slot

18. Hoe plan je een roadmap? Wat zijn de Best Practices en zaken waar we op moeten letten?
    Klein beginnen, probeer kleine use case die je van end – to – end kunt uitwerken.
    gegevens binnenhalen, gegevens verwerken tot einde gaan.
    Zien dat je een use case kunt uitwerken in begin.
    Heelvroeg in traject kan je bepalen of het haalbaar is.
    Klein beginnen – Business case maken (verkopen aan bedrijf) – verder opbouwen – niet alles te groot beginnen zien vanaf het begin. Hierdoor kun je snel op de bal spelen.
19. Zijn er nog zaken die wij niet hebben aangehaald hebben die belangrijk zijn om te vermelden.*Alles aangehaald. Belangrijkste zaken zijn gezegd.Klein beginnen , developers zijn belangrijk tijdens het proces, oppassen met tools zonder te weten wat er achter zit, goed team samenstellen (technische / data scientist).*
20. Zijn er specifieke bronnen over Big Data Analytics die u ons aanraadt die ons nog zouden kunnen helpen?
    linkedin, big data groepen, hadoop groepen, spark groepen, …
    Hadoopsumit / sparksumit ; Bellen als we vragen hebben.

9.1.4 Interview Deloitte

Algemeen

1. Hoelang bent u al bezig met Big Data of Business Analytics?

- Business analytics: sinds 1997, Big data : sinds 2012

2. Wat hebt u gestudeerd of hoe bent u bij Analytics uitgekomen?

- Arbeidssociologie met optie survey analytics
- Ik stuur mijn CV mee dan heb je zicht op mijn historiek
    - Survey bij TNS Sofres / Dimarso
    - Analytics met SAS bij Federale politie
    - Beheer van structured en unstructured data en business intelligence bij Flanders Investment & Trade
    - Business intelligence consultant bij Numius
    - Analytics & Information Management Consultant bij Deloitte

3. Zijn er genoeg mensen in België die zich bezig houden met Analytics?

- Nee, ik beschouw het als een knelpuntberoep en verwacht dat dit de komende jaren nog verder zal toenemen.

4. Is BA / DS echt iets voor business mensen of eerder voor meer IT mensen?

- Big data en business analytics lukt pas als typische business skills en IT skills goed gecombineerd worden. Daarnaast zijn ook vaardigheden als story-telling belangrijk.
- Basis is dat mensen die afstuderen met bepaalde skills over de interesse en nieuwsgierigheid moeten beschikken om gedurende hun carrière bij te leren in andere vaardigheden.
- Mensen met business skills moeten bijleren over wat er met technologie mogelijk is. Ze moeten het nut en de mogelijkheden snappen en zorgen dat er aan IT mensen de juiste dingen gevraagd worden.
- Mensen met IT skills moeten dan weer een gezonde passie hebben om te begrijpen waar het in de business om draait.
- In de markt is er een grote vraag naar personen die beide kanten kunnen begrijpen. Mensen die “spreekbuis” zijn of “vertaalwoordenboek” tussen de typische vaardigheden van IT en de inzichten van business zijn in analytics nog meer dan ooit aan de orde.
- Binnen analytics gaan beide skills hand in hand.
- Niemand is gebaat met Big data. Iedereen is op zoek naar Big Impact. Als de investeringen in technologie groter worden (Hadoop clusters, big data appliances, beheer van een groter park aan data assets) en er meer en meer data (makkelijker) voor handen is, moet er nog beter gekozen en gefilterd worden. Een DWH aanleggen met alle data assets is misschien niet zinvol als maar x% van die data ooit zal benut worden. Voor elke aparte vraag opnieuw data gaan cleanen en storen in een handig formaat is ook niet aan te raden.
- Het is meer en meer een uitdaging om de goede balans te vinden tussen automatiseren en ad hoc vragen voor inzichten beantwoorden. Dit vergt een goed begrip van business uitdagingen en creativiteit in de manier waarop IT automatiseert.

5. **Als je een data scientist zou moeten beschrijven in 3 woorden, welke zouden deze dan zijn?**

- **Nieuwsgierig** naar inzichten die bruikbaar zijn voor beslissingen (geen nice to know maar inzichten waar beslissingnemers in organisaties iets mee zijn)
- Kan **goochelen** met data en algoritmes
- Kan op een **sterk visuele of storytelling** manier inzicht overbrengen aan een groter publiek beslissingnemers

6. **Is Big Data en Data Analytics een hype of eerder nog steeds iets veel belovend?**

- Het zijn buzz woorden waarvoor nog veel organisaties aan het zoeken zijn op welke manier ze het veel belovend kunnen maken voor hun organisatie.
    - Het is niet genoeg om te investeren in technologie, data en competenties

        - een Hadoop cluster of allerhande analytical oplossingen
        - data scientists die kunnen algoritmes ontwerpen
        - aankopen van relevante datasets
    - Het is wel belangrijk dat de inzichten uit big data en de mogelijke manieren om ze te verwerken zinvol zijn.
        - Inzicht bepaalt het gedrag van mensen. (bv. recommender oplossingen op webshops, Waze data die mij 'smorgens helpt files vermijden)
        - Inzicht helpt organisaties keuzes maken (bv. al dan niet toekennen van lening op basis van risico modellen)
        - Inzicht helpt risico's opvangen (bv. predictive maintenance)

Opgelet met definitie van big data (grote volumes, snelheid, unstructured, ...) De typische V's van big data zorgen elk voor andere uitdagingen en mogelijkheden. We zien ook dat intensieve analyse (algoritmes) op relatief kleine hoeveelheden ook voor uitdagingen zorgen.

Je kan "Big" ook interpreteren als "moeilijk, uitdagender, nieuwe mogelijkheden" dan wat organisaties gewoon zijn. We zien organisaties die door de impuls van big data beslissen om bestaande processen te optimaliseren (OPTIMISE), anderen gaan dan weer hun manier van werken (REDEFINE), tot slot zijn er organisaties die door impuls van deze technologie compleet het roer omgooien (DISRUPT).

7. In welke onderwerpen/opdrachtsoort hebt u het meeste ervaring?

Begeleiden van organisaties in de transitie naar een beter gebruik van data en analytics. Helpen analyseren waarom ze de verwachte meerwaarde niet realiseren. Trachten de ROI te verbeteren door betere keuzes qua investering, change management in dit kader (bv. digitale transformatie)Ik heb verder heel wat ervaring in bouwen van business intelligence en performance management oplossingen.

8. Met welk platform hebt u het meeste ervaring mee of raadt u het meest aan (Big Data)?

Technisch platform is zelden de motor van succes of bron van falen. Mijn ervaring is dat de mogelijkheden en beperkingen van platformen of oplossingen goed kennen en transparant maken aan alle betrokkenen in projecten (zowel business als IT) van groot belang is.Elk platform of oplossing heeft zijn sterktes en zwaktes. De specifieke case maakt dat de keuze voor een oplossing op een gegeven moment altijd kan gemaakt worden. De meeste vendors vernieuwen continu hun technologie.Ik zou in het huidig klimaat vooral rekening houden met de mate waarin oplossingen open zijn naar integratie met andere platformen en oplossingen van andere vendors.We zien ook de grote vendors (SAS, IBM, Oracle, ...) steeds meer de openheid en integratie met open source oplossingen bepleiten en mogelijk maken. De kracht ligt in integratie en openheid en dat hebben alle visionairen in deze software markt al heel goed door.

9. Wat is belangrijk bij het verzamelen van data?

Data verzamelen op zich heeft geen nut. Men vraagt zich beter af welke beslissingen doorgaans moeten genomen worden en welke informatie deze beslissing kan verrijken. Als je dan toch data wil verzamelen zijn de volgende zaken belangrijk:

- Zorg voor goede metadata zodat later gebruik makkelijker is.
- Zorg voor zoveel mogelijk informatie over plaats en tijd waar data gecapteerd werd (al te vaak vergeet men dat data niet goed in de tijd plaatsen, riskeert dat foute conclusies genomen worden)
- Data en informatie management is een domein met zeer veel taken waarin security en privacy ook steeds belangrijker worden.
- Het besef dat data verzamelen, stockeren en beheren in een organisatie processen en dedicated mensen vereist en geen puur technische aangelegenheid is.

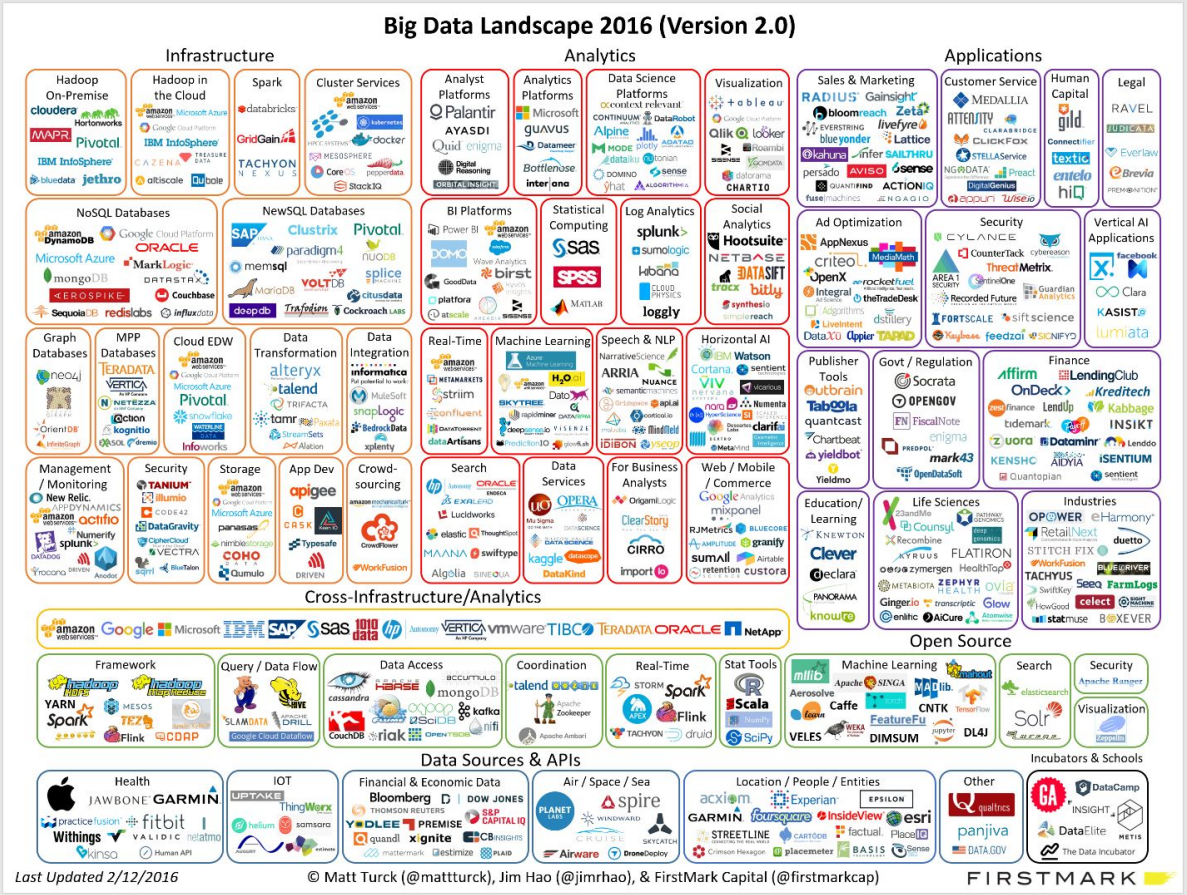
10. Wat zijn de belangrijkste use cases (Big Data)?

Er zijn er veel. Ik zie de laatste tijd meer en meer big data opduiken binnen Deloitte en onze klanten in het kader van :

- Sensor data voor predictive maintenance
- Online clickstream en recommender systemen (“anderen kochten ook, dit zijn de voor u meest interessante producten, vacatures, ...)
- Analyse van verkeersstromen (Waze, Oyster card Transport for London, ...)
- Camerabeelden en automatic number plate recognition
- ...

11. Wat zijn de belangrijkste Big Data Technologieën?

Ik vind onderstaand schema wel een goed schema. Ik vermeld hier en daar technologieën die ik in mijn activiteiten vaak zie opduiken. Ik heb de data bronnen en applicaties even niet bekeken.



Infrastructuur met Hadoop (cloud en on premise) en enterprise ready versies (Cloudera, IBM Biginsights, ..) NoSQL en NewSQL databases (MS Azure, MongoDB, ...) Graph databases (Neo4j, OrientDB, ...) Massive Parallel Processing databases (Teradata, IBM PureData for analytics (Netezza), Vertica) Hadoop world (Spark, Yarn, Hive, Mapreduce, Pig.) Analytics (SAS, SPSS, R, IBM Watson, Matlab, Python, ...) Visualisatie (Tableau, QlikSense, QlikView, D3) Big vendors: SAP (Hana), IBM, SAS, Oracle, Microsoft, Teradata,

12. Hoe moet een organisatie zich aanpassen voor Big Data?

- Vooral veel energie steken in dialoog tussen verschillende disciplines (business en IT) om de meest zinvolle toepassingen te selecteren.
- Kiezen voor een aanpak waarin meerdere mogelijkheden in een innovatie-lab uitgetest worden en vervolgens de beste cases verder worden uitgerold. Dit is vooral een trial en error, agile aanpak die het beste werkt voor dit domein.

- Zorgen dat de juiste mensen in de organisatie moeten samenwerken en zorgen dat er specifieke profielen (functies) zijn die met deze topic bezig zijn. Verwacht niet dat het morgen zonder effort waarde zal opleveren.

13. Welke invloed heeft Big Data op een bedrijf?

Big data kan zorgen voor een optimalisatie van bestaande processen. Het kan zorgen dat bedrijven zich moeten herorganiseren omdat ze dreigen hun markt volledig te verliezen of het kan ervoor zorgen dat de markt in een totale disruptie terecht komt. Mijn ervaring met innovatieve technologie is dat wendbare bedrijven zullen overleven en dat bedrijven die het moeilijk hebben met verandering en wijziging het lastig zullen krijgen. Vbn.

- Denk aan een taxibedrijf bij ontstaan en de mogelijkheden van Uber.
- Denk aan tomtom als iedereen via een app waze zicht krijgt op verkeersstromen en mogelijke alternatieve wegen
- Denk aan auto verzekeraars die via slimme wagens zicht zouden kunnen krijgen op het rijgedrag van chauffeurs om zodoende te diversifiëren in de tarieven voor een autoverzekering (voorzichtige chauffeurs betalen minder)

14. Met welk platform hebt u het meeste ervaring mee of raadt u het meest aan (Analytics)?

Beste oplossing is verschillend naargelang specifieke situatie of use case.

- Bedenk dat oplossingen best integreren met de bestaande architectuur bij de klant. De openheid van oplossingen voor integratie met andere oplossingen is daarom sterk aan belang aan het winnen. Het landschap wordt meer versnipperd tussen verschillende grote vendors en best of breed oplossingen.
- Het licentie model van de software verkoper mapt al dan niet met gewenste aantal gebruikers (aantal gebruikers is gekend en beperkt vs het aantal gebruikers is identiek aan alle burgers).
- De mate waarin oplossingen business kritisch zijn vergen de juist support organisatie wat door grotere vendors kan afgedekt worden en mogelijks door open source vendors niet.
- Bovendien veranderen oplossingen continu. De meeste organisaties kunnen zich niet permitteren om snel van oplossing te veranderen. Vaak is daarom ook de verwachte uitbreidingen of de investeringen die software vendors de komende jaren zullen maken van belang.
- Elke oplossing is een vat aan verschillende functionaliteiten. Goed weten wat je wil en nodig hebt, blijft aan de orde.
- Vergelijk dit met de aankoop van een smartphone of digitale televisie. Er zij opties in alle prijscategorieën en beste koop is afhankelijk van wensen aankoper.

15. Wat zijn de belangrijkste zaken om op te letten om een analytics platform op te stellen?

Ik zou durven zeggen: mate van openheid naar andere oplossingen en mate van ondersteuning (externe consultants, technische ondersteuning van vendor zelf) in de Belgische markt. We zien vaak dat klanten vooral kiezen voor oplossingen en dat ook de jobmarkt met kennis over oplossingen (bij zusterbedrijven, bij consultants, in scholen, …) een belangrijke indicator is om succesvol zaken te kunnen bouwen. Niemand is gebaat met een oplossing die in theorie goed is maar een paar specialisten vergt uit een ver land om ze aan de praat te krijgen.

16. Wat zijn de belangrijkste use cases (Analytics)?

Ik zie nog altijd veel segmentatie, clustering, monte carlo simulaties, optimalisatie toepassingen. Daarnaast binnen cognitieve chatbots meer en meer op te komen.

17. Welke invloed heeft Analytics op een bedrijf?

Meer en meer nodig omdat alle verbanden tussen relevante data handmatig onderzoeken niet meer haalbaar is. Meer en meer nood om relevante patronen op te sporen. Zelfde type aan impact op bedrijven. Optimize, redefine, disrupt. Ik merk dat de druk in de markt groter geworden is.

18. Hoe plan je een roadmap? Wat zijn de Best Practices en zaken waar we op moeten letten?

Hiervoor verwijs ik graag naar het rapport Big data for policy making dat ik jullie bezorgde.

Denk dat een roadmap best rekening houdt met de volgende assen. We merken dat éénzijdig op één van deze topics inzetten meestal niet de gewenste resultaten oplevert.

- Strategie: wat willen we bereiken? Hoe kan big data en analytics hierbij helpen? Prioriteiten
- Mensen: wie moet hiervoor samenwerken? Wie laten we dit trekken en beslissen? Welke skills hebben we meer en meer nodig en hoe bouwen we die op (uitbesteden, aanwerven, trainen)
- Technologie: Is onze enterprise architectuur aangepast aan de noden van morgen? Integreert alles waar nodig? Zijn onze oplossingen perfomant? Waarvoor gebruiken we cloud vs on premise?
- Data management: kwaliteit, governance, privacy, security, metadata, master data
- Processen: hoe zorgen we dat belangrijke processen goed lopen? (definitie van behoeften, keuzes van prioriteiten, agile vs waterval methode voor ontwikkeling, …)

19. **Zijn er nog zaken die wij niet hebben aangehaald hebben die belangrijk zijn om te vermelden.**

Big data en analytics is een domein dat aan belang wint. We merken dat succes vergt dat mensen brede skills hebben (purple people) maar tegelijkertijd worden de mogelijkheden en uitdagingen moeilijker waardoor verschillende disciplines opduiken en het onmogelijk is om alles te beheersen. Het is daarom belangrijk dat mensen beseffen dat meer en meer samenwerking tussen mensen met verschillende specialiteiten aan de orde is. Iedereen wordt verwacht continu bij te leren. Daarom is het in een opleiding vooral belangrijk om vaardigheden te kweken om snel nieuwe dingen te leren (vermogen om bij nieuwe zaken snel te googlen en bij te leren).

20. **Zijn er specifieke bronnen over Big Data Analytics die u ons aanraadt die ons nog zouden kunnen helpen?**

Recente projecten en verwezenlijkingen van Deloitte in de big data en analytics wereld

Studie voor de EU

https://joinup.ec.europa.eu/sites/default/files/dg_digit_study_big_data_analytics_for_policy_making.pdf

Voor VDAB – oplossing op basis van IBM Watson

https://www.youtube.com/watch?v=VaUoUHclFZk (filmpje gemaakt voor World of Watson in Las Vegas vorig jaar)

https://vimeo.com/168823228 (testcase in VDAB jobbar – de robot is uiteraard maar een gimmick die het leuker maakt voor de gebruiker maar wordt in een chatbot vervangen door schriftelijke tekst)

Voor Tele Ticket Services

https://www.youtube.com/watch?v=-VxTRrqXiTQ

Voor Europese Commissie –Humanitaire hulp

https://www.youtube.com/watch?v=zX1GqFGgE8w

Ook in minder klassieke omgevingen als HR data begint dit op te duiken. Zie hieronder ook een recente overname die we deden in dit domein.

https://www2.deloitte.com/be/en/pages/about-deloitte/articles/Deloitte_acquires_iNostix.html

9.1.5 Interview SAS

Algemeen

1. Hoelang bent u al bezig met Big Data of Business Analytics?
   Toon: 11 jaar BI (8 jaar sas) – pre sales (naar klanten/prospecten gaan voor demo's) voor wat is relevant en welke waarden kunnen klanten er uit halen (meer technologisch : data management Sas platform)
   Tine: rond de 10 jaar ervaring BI / analytics. Strategieën uittekenen. Hoe werkt de klant vandaag en hoe gestructureerde manier laten verbeteren.
   Elisabeth: 21 jaar – verschillende dingen op sas; en nu academische manager
2. Wat hebt u gestudeerd of hoe bent u bij Analytics uitgekomen?
   Toon: Burgerlijke ingenieur elektronica + master bedrijfseconomie. Gewerkt bij materialize (prototyping / medische implantaten), daarna 3 jaar als consultent (BI / budget en planning)
   Tine: Master Wiskunde + beleidsinformatica. Functioneel analist bank (DWH). Na 4 jaar bij SAS (na paar jaar rapporteringszaken OLAP / …) na jaar of 5 naar algemene rol (sales support maar niet voor bepaalde tech/domein) wat is de waarde dat big data (analytics) kan bieden rond bedrijven.
   Elisabeth: Master in geschiedenis, sinds 1996 sas, doorgegroeid naar training en les geven in SAS (tot 2003) Daarna interne projectmanagement project opgestart, 2 jaar projectmanagement, sinds 2013 academic manager.
3. Zijn er genoeg mensen in België die zich bezig houden met Analytics?
   Nee, te weinig mensen met SAS en te weinig analytics
   Er zijn complete Data scientist te weinig.
   Bedrijven proberen het op andere manier op te lossen. B mensen die via point en klik toepassingen analysen gaan toepassen.
   Dat gaat niet zo snel veranderen. Er zijn niet veel wiskundige of mensen met focus op statistiek en niet iedereen wordt Data scientist.
   wordt mede opgelost door team verband werken.
   Mensen van externe partij tijdelijk inhuren.
4. Is BA / DS echt iets voor business mensen of eerder voor meer IT mensen?
   De pur sang data scientist moet affiniteit hebben met business en geïnteresseerd zijn in IT en probleemoplossend denken (door middel van coderen). Meestal komen data scientist uit business. Sommige bedrijven kiezen ze om bij IT te plaatsen. Soms onder Strategie departement.
   Zijn verschillende modellen (data scientist in centraal team, virtueel team)
   in realiteit: veel data scientist alleen werken : gevaar is schitterend werk dat niet gebruikt wordt. ZE slagen er niet in om business mensen te overtuigen.
5. Als je een data scientist zou moeten beschrijven in 3 woorden, welke zouden deze dan zijn?
   Flexibel, open minded, eager to learn

6. Is Big Data en Data Analytics een hype of eerder nog steeds iets veel belovend?
   Het is over de hype heen (volgens Gartner en It). Het is er , we zijn voorbij de periode dat mensen zeggen dat we er iets mee moeten doen, en wordt gewoon gebruikt.
   vanuit sas: maakt niet uit of het big data of data aan hoge snelheid, we moeten er gewoon mee om gaan.
   Maakt niet uit; data is data en als je er waarde wil uithalen moet je er iets mee doen. Analytics en er iets mee doen.
   Binnen big data zijn er veel zaken die eventjes hype worden. (Machine Learning)
   Die gebaseerd zijn op iets dat lang bestaat.
   analytics is op dit moment belangrijk voor bedrijven om competitief voordeel mee uit te halen en bewaken. Analytics om zich te differentiëren.

7. In welke onderwerpen/opdrachtsoort hebt u het meeste ervaring?
   zie vraag 10

Big Data

8. Met welk platform hebt u het meeste ervaring mee of raadt u het meest aan (Big Data)?
   Best onderscheidt tussen storage bronne (fysieke opslag) en data processing
   storage: Oracle / SQL / Hadoop; meer ingang: Hadoop eco systeem (beginpunt van echt in gebruik nemen) Dit linkt met data lakes (wat een hype is) => Beginnen ze in vraag te stellen of ze hun beloftes in vraag stellen
   theoretisch / archecturale : lake / hadoop inladen zonder structuur
   Data lake: data zetten zodat het is, niet integreren, cleanen, … wie het gaat gebruiken moet het maar opkuisen (tegenstrijdig met business mensen die dit willen gebruiken en niet kennen : Beperkingen theorie vs. praktijk)
   praktische: niemand gebruikt data lake in die divisi.
   Meer in labs gebruiken ze datalake om stromen te stokeren.
   Data processing: SAS
   gemakkelijk maken voor niet technische mensen om met hadoop te werken.
   hbase / hive heel veel code.
   Sas interfaces point en klik boven hadoop / spark omgeving bouwen.
   Makkelijk voor data management.
   transponeren voor ideale gegevens.

9. Wat is belangrijk bij het verzamelen van data?
   context maakt veel verschil uit.
   bronnen, toegankelijkheid, aan data geraken, automatiseren, snappen wat in bronnen steekt. metadata (hoe zit het er in , hoe ziet het er uit, business context => MASTER DATA MANAGEMENT) tools om beheren, Qchecks, matching, juiste regeles/sleutels , dictionairy. Niet veel bedrijven gaan zover. Moeilijk voor nieuwe mensen.
   Betrouwbaar zijn, goede kwaliteit. Manieren voor gemakkelijke : Virtuele laag boven plaatsen. Eindgebruikers van bronsystemen afschermen. Waar zit de data ? (DWH / Lake) en data samenbrengen. Identificeren van data (niet altijd evident – zit in veel systemen). documenteren van data en linken. klanten stellen meer en meer de vraag of een Enterprise Database Management nodig(DWH, Ster, …)
   Heel vaak zien: spanningsveld tussen business en IT (business ziet geen added value in IT) tools moeten faciliteren dat business en it beter met elkaar werken.
   vanuit sas: tools op gecontroleerde manier aanbieden aan business.
   tools moeten meer waarde bieden voor IT en business.
   bi portal: catalogue van rapporten.
10. Wat zijn de belangrijkste use cases (Big Data)?
    Marketing (klant beter leren kennen , customer 360 – turn verminderen, meer laten afnemen, recommedations) Retail (alles wat te maken heeft met verkoop / iot – klant loopt in winkel met app (praat met ibeacons en krijgt perso. boodschap)
    predictive: voorspellen wanneer machines defect gaan zijn (barco projectoren, voorspellen wanneer lampen kapot gaan) IoT / sensordata : manufactering
    Fraude en security: fraude detectie (HOT) bij banken/verzekeraars/ FOD Financie)

Bij belastinginspectie btw carrousels om belastingfraude tegengaan (met SAS)

Data met smartmeters fraude detecteren rond zonnepanelen. Meer elektriciteitsgebruik dan aangegeven.

11. Wat zijn de belangrijkste Big Data Technologieën?
   SAS en Hadoop (storage en processing platform)
   SAP Hana, Terradata, Oracle Big Data appliance
   Cloudera/Hortonworks -> Vaak gepackaged door andere (appliance leveren op basis van …

)
Vanuit Sas: puur databron (bron in sas) of in data store processing (rekenwerk naar appliance naar Hadoop pushen)

12. Hoe moet een organisatie zich aanpassen voor Big Data / analytics?
    Niet me big bang, ervaring met materie op doen, meer experimenteren, POC maken, daaruit leren en groeien ; uiteindelijke doel: omgeving voor operationeel en omgeving voor te "spelen" . Dingen in experimenteeromgeving die waardevol zijn moeten geïndustrialiseerd worden. Andere manier van werken: waterfall werkt NIET, eerder agile approach Sprints werken, geraken we ergens ?
    Business moet de tools hebben.
    Sas tools vs. IT tools
    ETL tools
    People: IT / kennis van systemen/ processen vs. Data Scientist (Citizen data scientist)
    Veel bedrijven hebben geen plaats voor (Citizen) data scientist
    Bedrijf moet inzicht hebben in analytics (management moet er achter staan)
13. Welke invloed heeft Big Data op een bedrijf?
    zie alle vragen

Big Data Analytics

14. Met welk platform hebt u het meeste ervaring mee of raadt u het meest aan (Analytics)?
    Sas, en zie vraag 11
15. Wat zijn de belangrijkste zaken om op te letten om een analytics platform op te stellen?
    Vermijdt dat je de 2 werelden kan bedienen, schaalbaar, flexibel (tekst mining relevant of niet), groei anticiperen
    juiste technologieën gebruiken
16. Wat zijn de belangrijkste use cases (Analytics)?
    zie vraag 10
17. Welke invloed heeft Analytics op een bedrijf?
    zei vraag 12

Slot

18. Hoe plan je een roadmap? Wat zijn de Best Practices en zaken waar we op moeten letten?
    Typisch gebeurd dat via brainstorm sessies waar er gezocht wordt voor relevante use cases die geprioriteerd worden waar dan een roadmap van gemaakt wordt.
    Change management is hier belangrijk in, automatiseren van zaken.
    Analytics kan hier bij helpen. Menselijke input mag je niet vergeten.
    Beste is analytics en menselijke input samenvoegen = Krachtige tool.
    Rond cases + stap voor stap (klein beginnen, met successen overtuigen, verantwoorden om grotere omgeving te bieden, schaalbaarheid, mensen overtuigen voor budgetten te krijgen)
19. Zijn er nog zaken die wij niet hebben aangehaald hebben die belangrijk zijn om te vermelden.
    /
20. Zijn er specifieke bronnen over Big Data Analytics die u ons aanraadt die ons nog zouden kunnen helpen?
    White papers op website SAS, Blogs, LinkedIn