Moduleopdracht

Ontwerpen en Programmeren

S. Crisan

Studentnummer: 4689887

Datum: 15-06-2019

NCOI

HBO Bachelor Informatica

Ontwerpen en Programmeren

Docent: Erik Mols

Voorwoord

Mijn naam is Sebastiaan Crisan. Van 1 november 2018 t/m 28 Maart 2019 heb ik stagegelopen als developer bij ADchieve in Rotterdam (hoofdkantoor Den Bosch). Ik ben begin september gestart met de opleiding “HBO Bachelor Informatica”. Met behulp van mijn stage heb ik een goede basis werkervaring opgedaan die ik hopelijk samen met de kennis uit mijn studie op het gebied van informatica kan toepassen en mij later te verdiepen.

Als stagiair developer was het mijn taak om een automatische CSS (Comparison Shopping Services) aanmelding van merchants met een Google merchant-id te realiseren via de website van de cliënt. Tevens werk ik aan een automatische reporting-tool die data vanuit Google Ads ophaalt, opslaat in een database en vervolgens aggregeert op basis van de business interests. Ik hoop met behulp van deze opdracht mijn kennis en inzichten te vergroten en deze later toe te kunnen passen bij mijn baan.

Samenvatting

Er is gekozen om de applicatie die ontwikkeld is voor ADchieve, te versimpelen en als Java-applicatie te schrijven in het kader van de moduleopdracht. Eerst zal het doel van de applicatie uitgelegd worden. Vervolgens zullen de functionaliteiten van de applicatie uitgelegd worden. Daarna zullen de klassen van de applicatie weergegeven worden door middel van een UML klassendiagram.

Inhoudsopgave

Inleiding

ADchieve is een snelgroeiend bedrijf, hoofdkantoor gevestigd in Den Bosch, met een tweede nieuwe locatie te Rotterdam. Het bedrijf richt zich op het maximaliseren van winst bij cliënten (merchants), die reclame willen maken via Google Ads en Google Shopping, met behulp van algoritmen en big data-

analyse. De moduleopdracht zal in gaan op het deel van de stageopdracht wat betrekking heeft tot het ophalen, opslaan, aggregeren en weergeven van data.

Het doel van de moduleopdracht is het inleveren van een verbeterplan voor het verbeteren van datamanagement bij ADchieve. Dit begint aan de hand van een SMART geformuleerd vraagstuk. Dit vraagstuk bevat: aanleiding, doelstelling en probleemstelling. Tevens zal beschreven worden hoe dit vraagstuk zich de laatste periode heeft ontwikkeld, wat de huidige stand van zaken is en wat hierbij de voornaamste knelpunten zijn. Vervolgens zal er gekeken worden naar wat de gewenste situatie is en op welke wijze deze situatie een oplossing biedt voor de daar voorafgaand geschetste problemen. Als laatst zal een conclusie geformuleerd worden met daarin aanbevelingen voor ADchieve, en de beantwoording van de probleemstelling. In de moduleopdracht komen minimaal de volgende onderwerpen aan bod:

- Een visie op datamanagement van gestructureerde en ongestructureerde data bij ADchieve
- De eisen aan data-kwaliteit
- De rollen en verantwoordelijkheden t.a.v. datamanagement bij ADchieve

Hoofdstuk 1: Probleemstelling

Datamanagement is een belangrijk onderdeel bij ADchieve. Data van klanten en van hun klanten wordt opgeslagen in de database. Het bedrijf groeit snel, en daarbij komt dus veel nieuwe functionaliteit kijken en daarbij nieuwe data- en informatiebehoeften. Met andere woorden: Het zou erg handig zijn als er een tool ontwikkeld zou worden die big data op een bepaalde manier aggregeert, op

basis van de huidige business behoeften. Bijvoorbeeld: Hoeveel hebben alle klanten in het vorige kwartaal aan ads gespendeerd, en op basis hiervan, wat zou de voorspelling zijn per klant voor de ad spend in dit kwartaal?

Min of meer gaat het hier om het ontwikkelen van een data warehouse. Een data warehouse is in de simpelste vorm een kopie van data, die op een bepaalde manier gearrangeerd is zodat deze gemakkelijk op te halen is. Deze data kan vervolgens op een bepaalde manier geaggregeerd worden, en hier kan dan waardevolle informatie uit gehaald worden, bijvoorbeeld voorspellingen van prestaties van een business (Gordon, 2013). Omdat ADchieve tot nu toe redelijk informeel te werk is gegaan wat data-management betreft, zijn hier veel verbeteringen in aan te brengen. Als basis voor deze verbeteringen is voor te stellen dit direct toe te passen op het ontwikkelen van deze tool (data warehouse). De probleemstelling is te formuleren aan de hand van de obstakels die zich voordoen tijdens de ontwikkeling van de tool met betrekking tot datamanagement, eisen aan data-kwaliteit en de rollen en verantwoordelijkheden t.a.v. datamanagement bij ADchieve:

Probleemstelling: Welke verbeteringen moeten worden doorgevoerd om datamanagement op een hoger volwassenheidsniveau te krijgen, zodat tegen het eind van de ontwikkeling van het proof-of-concept van de nieuwe tool er formele eisen aan data kwaliteit, duidelijke rollen en verantwoordelijkheden t.a.v. datamanagement, en een duidelijke visie op datamanagement van gestructureerde en ongestructureerde data bij ADchieve zijn?

Hoofdstuk 2: Huidige Situatie en Knelpunten

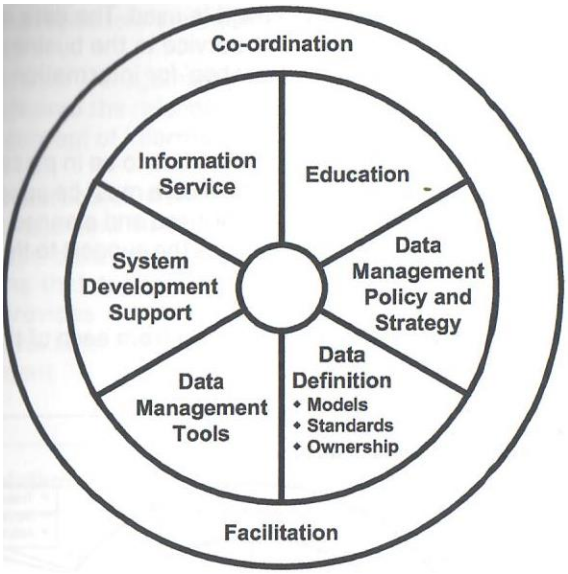
Allereerst is het handig om te beschrijven wat de huidige situatie bij ADchieve is, hoe het er precies aan toe gaat in bepaalde gebieden, met name:

- Wat zijn de rollen en verantwoordelijkheden m.b.t. datamanagement?
- Wat zijn de eisen aan data-kwaliteit?
- Wat is de visie m.b.t. gestructureerde en ongestructureerde data?

Na het beantwoorden van deze vragen zal blijken wat de knelpunten en verbeterpunten zijn bij de ontwikkeling van een formelere vorm van datamanagement.

Rollen en verantwoordelijkheden m.b.t. datamanagement
Definitie van datamanagement: “Datamanagement is een bedrijfsservice die helpt bij het voorzien van informatie-services, door middel van het besturen en coördineren van definities en het gebruik van betrouwbare en relevante data” (Gordon, 2013)

De rollen en verantwoordelijkheden m.b.t. datamanagement is eenvoudig in kaart te brengen door datamanagement-activiteiten weer te geven door middel van het volgende figuur:



Figuur 1 – Datamanagement-activiteiten (Gordon, 2013)

Het figuur heeft zes verschillende schijven, elke schijf een categorie van activiteiten in het datamanagement. Elke activiteit heeft bepaalde deliverables.

Datamanagement-activiteiten:

- *Education*
  - Trainingen (bijvoorbeeld Google Ads Scripts course in Londen, of trainingen gerelateerd aan de manier van werken in het bedrijf zelf)
  - Afstudeerprojecten ondersteuning
  - Stage
  - Lunch & Learn Sessions (Collega’s informeren over huidige activiteiten binnenin het bedrijf met als doel dat iedereen op een informele manier een goed beeld krijgt van de activiteiten en structuur binnenin het bedrijf)
- *Data Management Policy and Strategy*
  - Jaarmeetingen
  - Overleg elke 2 weken
  - Er is geen formeel beleid gedocumenteerd m.b.t. de datamanagement-strategie
  - Scrum meetingen

- *Data Definition (models, standards, ownership)*
  - Voorgeschreven door tools (klantdata, wat is een ADchieve account, ads data)
  - Scrum proces (eigenschappen van een item: definition of ready, definition of done, testen, nieuwe functionaliteit)

- *Data Management Tools (tool die men in staat stelt data(processen) te ordenen, bekijken, wijzigen, verplaatsen en op te slaan met als doel hier in de toekomst informatie uit te halen)*
    - MySQL
    - Sequel PRO
    - Pipedrive (CRM, handelt veel klantdata)
    - Online harde schijf
    - ADchieve zelf

- *System Development Support*
    - Product Owner die nieuwe functionaliteit uitwerkt met requirements en modellen

- *Information Service*
    - Wiki pagina over ADchieve op GitHub
    - Handleiding geschreven door één van de werknemers
    - Frontend documentatie gegenereerd uit code

Door deze activiteiten te analyseren zijn er verbeterpunten uit te halen. Het is goed dat er trainingen gegeven worden waar daar nodig is, of om op te frissen. Wat beter kan is het feit dat er geen formeel beleid is gedocumenteerd m.b.t. de datamanagement-strategie. Hier kan dus verbetering in aangebracht worden door een datamanagement-strategie te formuleren en te documenteren. Er zijn voldoende datamanagement-tools aanwezig, en binnenkort ook de nieuwe tool die ontwikkeld gaat worden. De Product Owner houdt zich actief bezig met het doorvoeren van ontwikkelingen qua functionaliteit binnen het bedrijf, en zorgt dat deze duidelijk zijn door modellen en requirements uit te werken. Er is informatie beschikbaar over het bedrijf in de vorm van documentatie van de code, een wiki-pagina op GitHub en een handleiding geschreven door werknemers. Wat hieraan ontbreekt is een handleiding voor formele data-definities, wat de structuur van de database is, en wat de input en output is. Hier kan dus ook verbetering in aangebracht worden. Er zijn verder geen echte knelpunten aanwezig, het is vrij eenvoudig de nieuwe verbeteringen aan te brengen omdat het vooral een kwestie van formeel documenteren is en goede afspraken maken.

De eisen aan data-kwaliteit
Nu de manier van werken duidelijk is, en welke verbeterpunten hierin aangebracht kunnen worden, is het mogelijk om te verdiepen en de eisen aan data-kwaliteit te analyseren. Wikipedia (2019) geeft aan dat data-kwaliteit volgens de ISO 9000:2015 “de mate waarin een set eigenschappen van data voldoet aan de eisen waarvoor deze gebruikt gaan worden”, is.

ADchieve verwerkt voornamelijk klantgegevens en gegevens die klanten invoeren in het systeem. (Klanten voeren allerlei instellingen in via UI (User Interface) van de software). Verder heeft elke klant ook zijn/haar eigen product feed, van waaruit data komt. Aan dit alles zijn indirect bepaalde kwaliteitsstandaarden verbonden:

- *Klantgegevens*
    - Geen *duplicates* (dit is belangrijk zodat bij een lead, er niet meerdere mensen van sales achter dezelfde lead aan zitten en ‘spam’ veroorzaken).
    - Moet up to date zijn. Een verkeerd telefoonnummer of adres heeft geen waarde.
    - Minder belangrijk, maar niet onbelangrijk: De data moet compleet zijn. Echter, heeft men in principe genoeg aan een klantnaam, adres en telefoonnummer.

- *Data die klanten invoeren in het systeem*
    - Deze data moet wel compleet zijn.
    - De data moet geldig zijn.

- *Product feed data*

- Hoeft niet foutloos te zijn (Een klant vergeet soms bijvoorbeeld een prijsveld in te vullen).
- Het hoeft ook niet compleet te zijn.
- De data wordt gevalideerd op geldigheid door een tussenlaag, voordat de data het systeem in gaat.

Er is veel te verbeteren t.a.v. de bovengenoemde richtlijnen. Zoals bleek bij de datamanagement-activiteiten, is dat er dingen niet formeel gedocumenteerd zijn, maar meer uit de losse pols geschud worden. Zo ook bij het documenteren van metadata. Metadata is data van data, bijvoorbeeld: 'Dit getal 1212 is het aantal reclamecampagnes die op dit moment actief zijn' (Gordon, 2013). Het probleem dat metadata niet formeel gedocumenteerd wordt, is wederom op te lossen door een formele definitie aan te brengen, dit formeel te documenteren en zich hieraan te houden bij alle data waarvoor dit is afgesproken. Dit geldt ook voor het documenteren van overige richtlijnen. Het is minder van belang dat de product feed data, vóórdat het de validatie-tussenlaag bereikt, correct is. Dit is een bewuste trade-off die gemaakt wordt tussen gemak van de klant en volledigheid. Overigens is de juistheid van de data van de klant, uiteindelijk de verantwoordelijkheid van de klant, en niet van ADchieve.

Een knelpunt wat zich nu voordoet, is omdat er geen formele definities zijn aangebracht voor metadata, dit dus extra kosten met zich mee gaat brengen in het kader van het implementeren van de reporting tool (data warehouse), omdat eerst (vage) data definities verzameld moeten worden voordat de reporting tool ontwikkeld kan worden. Het betreft nu slechts één tool, dus de impact is minimaal. Echter, zou het meerdere nieuwe implementaties betreffen, dan zou dit herhaaldelijk moeten gebeuren. Het zou efficiënter zijn als er voorgaand duidelijk de data definities gedocumenteerd zouden worden, als investering in opvolgende implementaties.

Visie m.b.t. gestructureerde en ongestructureerde data
In principe is alle data die ADchieve gebruikt, gestructureerd. Gestructureerde data is volgens Gordon (2013) in feite data die bestaat uit korte strings (bijvoorbeeld namen), nummers en datums.

Gordon (2013) heeft het ook over ongestructureerde data, ook wel multimedia-data. Dit kunnen plaatjes, grafieken, audio en video zijn. Er wordt binnen ADchieve wel gebruik gemaakt van informatie in de vorm van grafieken. Bijvoorbeeld bij de reporting-tool zal ook een aantal charts/grafieken zitten. Dit is echter informatie en geen data, de grafieken worden slechts gegenereerd en weergegeven om af te lezen op basis van geaggregeerde data. Ze worden niet opgeslagen in de database. Het zou echter wel verstandig zijn definities op te stellen voor ongestructureerde data, voor het geval hier in de toekomst nog gebruik van gemaakt gaat worden.

De toegevoegde waarde van het definiëren van data-elementen
Wanneer een data-element gedefinieerd is, kan men snel bepaalde informatie over de data achterhalen. Ook helpt het developers ervoor te zorgen dat deze data begrepen wordt door elk afzonderlijk systeem. Tevens is het van belang om te weten wat voor operaties er op deze data uitgevoerd kunnen worden, bijvoorbeeld het getal 12. Deze data is gedefinieerd als een currency, specifiek de euro, en als een type float met twee decimalen. Hierbij kunnen nummers worden opgeteld, afgetrokken, gedeeld of vermenigvuldigd. Echter niet, wanneer deze nummers een currency bevatten die niet dezelfde currency is, omdat deze dan eerst omgerekend moeten worden. Zonder data definitie was deze berekening in de soep gelopen (Gordon, 2013).

Data definities bij ADchieve
De data wordt op een database opgeslagen met data definities, en er is een handleiding voor het maken van data definities. Op deze manier wordt ervoor gezorgd dat alle data op correcte wijze gedefinieerd wordt.

Voornamelijk zijn er voordelen aanwezig, namelijk dat er geen foute eenheden uit de data voortkomen en tevens weet elk systeem dat dezelfde data definities hanteert met deze data om te gaan. Elk systeem hanteert dezelfde data definitie regels. Volgens Gordon (2013) is het dan goed geïntegreerd.

Nadeel kan zijn dat wanneer er een nieuwe data definitie bij komt, de handleiding bij geschreven moet worden en vervolgens op een globale scope de data definities bijgewerkt moeten worden wat dus voor ietwat extra werk zorgt, maar dit weegt veelal niet op tegen de moeite die er bespaard wordt. Over het algemeen kan data efficiënt verwerkt worden waardoor hier betere informatie uit gehaald kan worden waardoor de kwaliteit stijgt. In het kader van data definities zijn er dus geen knelpunten aanwezig.

Knelpunten samengevat
Na grondige analyse van de situatie bij ADchieve zijn dit de voornaamste knelpunten:

- Er is geen formeel beleid gedocumenteerd m.b.t. de datamanagement-strategie. Dit zorgt ervoor dat er niet optimaal en daarmee efficiënt gewerkt wordt.
- Er is geen formele definitie van metadata (dingen worden een beetje uit de losse pols. gedaan), dit is een risico m.b.t. de communicatie van databases, voor implementatie van nieuwe onderdelen gaat dit veel tijd kosten om te controleren en informatie bij elkaar te schrapen.
- Er gaat tijd in zitten om een beleid toe te voegen m.b.t. ongestructureerde data.

Hoofdstuk 3: Gewenste Situatie

Na het analyseren van de huidige situatie, zijn verbeterpunten/knelpunten verzameld. Op basis hiervan kan de gewenste situatie voorgesteld worden, waarin deze problemen opgelost zijn en ADchieve een efficiënter bedrijf is geworden.

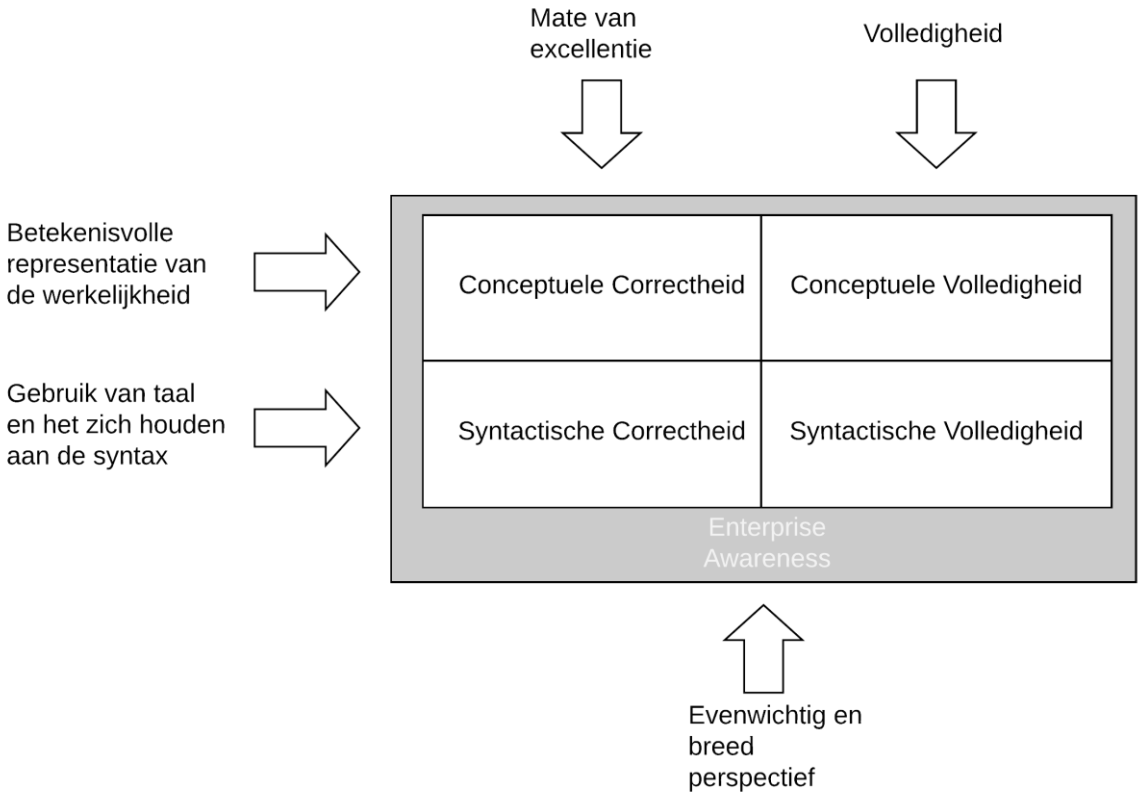
De gewenste situatie is er een waarin ADchieve globale en omvattende data-definities hanteert, zoals voor metadata, en formeel te werk gaat en documenteert. Dit houdt ook in dat er een formeel beleid is ingesteld m.b.t. de datamanagement-strategie. Dit lost de knelpunten/problemen op die ervoor zorgen

dat er elke keer opnieuw informatie verzameld moet worden over data definities, metadata, en dat data vaak gecontroleerd moet worden, op het moment dat er nieuwe functionaliteit geïmplementeerd gaat worden. Ook neemt het risico's weg m.b.t. de onderlinge communicatie van de databases.

Verder zal er gebruik gemaakt worden van het data warehouse, de reporting tool. Dit is mogelijk gemaakt doordat alle informatie beschikbaar is, en omdat er tegen die tijd dan formele kwaliteitsstandaarden zullen zijn gedocumenteerd, zal de data die de reporting tool gaat gebruiken ook zeker van hoge kwaliteit zijn.

De kwaliteitsstandaarden zullen minimaal voldoen aan het wegnemen van de door Gordon (2013) genoemde risico's. Een database zal een goede structuur hebben waardoor tijdens het overdragen van data van de ene database naar de andere, de data altijd op dezelfde manier geïnterpreteerd zal worden. Er zal meerdere malen gecheckt en gevalideerd worden door zowel mensen als systemen, of de input juist en geldig is, en of een geen *duplicates* aanwezig zijn. Er zal ook gecontroleerd worden of data niet gecorrumpeerd raakt tijdens het verplaatsen van het ene systeem naar het andere. Als laatste zal iedereen die met de data te maken heeft binnen ADchieve, goed geïnformeerd zijn over wat deze data betekent, zodat er geen misverstanden kunnen ontstaan m.b.t. de interpretatie van data en de informatie die daaruit gehaald wordt. Op deze manier vormt ADchieve een breed en evenwichtig perspectief m.b.t. de data, en haalt daar betekenis uit. Dit alles bij elkaar zal resulteren in data die correct en volledig zal zijn.

Hierdoor stijgt als het ware het bewustzijnsniveau van de business (Enterprise Awareness). Dit wordt ook schematisch weergegeven door Gordon (2013), in een model dat is ontwikkeld door Michael Reingruber en William Gregory. Zie 'Figuur 2 - vijf dimensies van data model kwaliteit' op de volgende pagina.



Figuur 2 - Vijf dimensies van data model kwaliteit (Gordon, 2013).

De vijf dimensies zijn voor de pijlen aan de buitenkant van het model weergegeven. Deze dimensies vertalen zich in vier categorieën, die samengevoegd de mate van het bewustzijnsniveau van de business betekenen.

Door deze data van hoge kwaliteit op een bepaalde manier te aggregeren, kan de Product Owner in één oogopslag zeer waardevolle informatie (Gordon, 2013) uit data van de afgelopen kwartalen halen, en eventueel voorspellingen maken voor aankomende kwartalen, en zal het algehele bewustzijnsniveau van ADchieve m.b.t. data sterk stijgen.

Hoofdstuk 4: Conclusie

Het is voornamelijk van belang dat er formeler te werk gegaan moet worden. Er is een handleiding met daarin data-definities voor gestructureerde data aanwezig, maar voor metadata is geen beleid. Er moeten dus duidelijke regels aan de handleiding toegevoegd worden voor metadata.

Voor ongestructureerde data is nog helemaal geen beleid aanwezig. Nu is het zo dat ADchieve hier op het moment geen gebruik van maakt, maar in de toekomst zou het kunnen dat dit gaat veranderen. Het loont dus als investering, hier ook definities en regels voor te maken.

Er is als laatste geen formeel beleid m.b.t. de datamanagement-strategie. Dit resulteert in een langzamere workflow, omdat elke keer opnieuw informatie over de manier waarop met data omgegaan dient te worden, verzameld moet worden. Dit kost tijd en daarmee geld. Ook hier zou het documenteren van een formeel beleid en zich hier vervolgens aan te houden, de efficiëntie flink omhooggooien. De data-kwaliteit moet voldoen aan de gedocumenteerde kwaliteitsstandaarden, en minimaal de risico's wegnemen die genoemd zijn door Gordon (2013). Het datamodel moet tevens van dusdanige kwaliteit zijn, dat wanneer de vijf dimensies van data model kwaliteit samengevoegd worden, het bewustzijnsniveau van ADchieve sterk stijgt waardoor er waardevolle(re) informatie verkregen zal kunnen worden uit bestaande en nieuwe functionaliteit.

Wanneer dit alles in orde is, kan nieuwe functionaliteit, zoals een data warehouse, efficiënt geïmplementeerd worden, omdat er formeel gedocumenteerd staat:

- Wat de datamanagement-strategie is
- Wat de data-definities zijn voor gestructureerde en ongestructureerde data
- Hoe metadata de data beschrijft
- Wat de kwaliteitsstandaarden zijn

Na het bereiken van deze doelen zal ADchieve met grotere zekerheid informatie uit data kunnen halen, omdat deze data van hogere kwaliteit is. Tevens zal ADchieve er hier als bedrijf veel bewuster en efficiënter op worden, en daarmee veel tijd en daarmee veel geld besparen waardoor er in de toekomst een grotere winst zal worden behaald.

Literatuurlijst

Gordon, K. (2013). *Principles of Data Management: Facilitating Information Sharing*. Swindon, United Kingdom: BCS Learning & Development Limited.

Wikipedia contributors. (2019, 19 maart). Data quality - Wikipedia. Geraadpleegd op 25 maart 2019, van https://en.wikipedia.org/wiki/Data_quality