Plan van Aanpak

Afstudeerstage, Efficiency-Online

Datum
september 2004
Auteur
Wouter Ewalds

Inhoud

1. Inleiding

In dit document zal beschreven worden hoe ik mijn afstudeerstage bij Efficiency-Online ga aanpakken. Deze afstudeerstage zal plaatsvinden in de periode september 2004 tot maart 2005. Dit document is een richtlijn voor het hele afstudeer proces. In dit document wordt zowel de opdracht als het proces besproken.

1.1 Efficiency-Online

Efficiency-Online levert producten en diensten op het gebied van CRM, financieel beheer, HRM, project management, kennismanagement en data analyse. Het bedrijf is gevestigd aan de Houtmankade in Amsterdam.

1.2 Begeleiding

Bij Efficiency-Online zal ik begeleid worden door Daniel van der Wallen, mede oprichter en directeur van het bedrijf. Zijn achtergrond ligt in de artificiële intelligentie en hij heeft de nodige jaren ervaring in de ICT industrie. Vanuit deze achtergrond zal hij mijn voortgang bewaken en me waar nodig ondersteuning bieden.
Vanuit de Universiteit zal ik begeleid worden door Patrick van Bommel. Hij zal zich voornamelijk bezig houden met de ondersteuning en bewaking van de academische kant van mijn stage.

1.3 Over dit document

In dit document zijn een aantal hoofdstukken opgenomen die een beeld moeten scheppen over de te nemen weg en de aanpak. Aangezien we ervoor gekozen hebben dit document als belangrijke input te laten dienen voor de toekomstige scriptie wordt er in deze fase een literatuurstudie gestart.

2. Opdracht

In dit hoofdstuk zal ik de opdracht die ik ga uitvoeren bespreken. Het doel van dit hoofdstuk is het afbakenen van de opdracht: wat hoort wel en niet bij mijn opdracht. Het doel van de stage is alles wat in dit hoofdstuk besproken wordt uit te voeren en te beschrijven in een scriptie.

2.1 Omgeving

Vrijwel iedere organisatie heeft tegenwoordig een website. Zeker de laatste jaren is het steeds belangrijker geworden om de gebruiker snel de relevante informatie aan te bieden. Dit proces kan gemeten worden door statistieken bij te houden en hierover uitspraken te doen. Deze case is een voorbeeld van wat je zou kunnen analyseren. Het onderzoek abstraheert hiervan.

Het analyseren zou met behulp van een *OLAP*[3], *O*nline *A*nalytical *P*rocessing, tool gedaan kunnen worden. De medewerkers van de organisatie zouden dit steeds vaker via het Internet willen doen, later aangeduid met *WOLAP*. Nu wil het feit dat *OLAP* een zware toepassing is en het *WWW*, *W*orld *W*ide *W*eb, in essentie een traag medium. Om verwarring te voorkomen; het woord 'Online' in de term *OLAP* betekent in dit geval niet hetzelfde als via het Internet 'Online' zijn. In *OLAP* betekent 'Online' de letterlijk vertaling, 'koppeling'. Het is via een koppeling met een 'centrale' computer en koppelingen met meerdere gegevens mogelijk om analyses uit te voeren. De toevoeging *W* aan *OLAP* wil dus zeggen dat je de bestaande techniek beschikbaar maakt voor het Internet (WWW).

2.2 De Opdracht

Globaal zou je de opdracht samen kunnen vatten in: "zorg ervoor dat gegevens snel kunnen worden geanalyseerd via het *WWW* zonder dat je van te voren weet wat een gebruiker wil weten." Nu is *OLAP* een uitstekende techniek hiervoor, alleen geeft een directe vertaling naar het web performance problemen. Hier ligt de kern van de opdracht. Om ervoor te zorgen dat gebruikers snel de informatie die ze zoeken kunnen bereiken, is het idee geboren om gebruikers profilering toe te passen. Dit gaat gebeuren via het Datamining tool Safarii, een software pakket van Efficiency-Online waar ik gebruik van mag maken.

Dit houdt concreet in:

- Het gaan beschikken over grote hoeveelheden data
- Het profileren van gebruikers
- Het transformeren van de output van het Safarii pakket naar een *OLAP* geschikte formaat
- Een analyse applicatie ontwikkelen die hiervan gebruik kan maken en zich aanpast aan de gebruiker
- Documentatie voor zowel gebruikers als systeembeheerders

Aan dit systeem, het geheel van alle componenten, zijn een aantal voorwaarden verbonden:

- De uiteindelijke analyse zal per gebruiker na vergaring van gegevens over deze gebruiker even snel of sneller moeten gaan dan de normale *OLAP* techniek. Dit is tevens een van de doelen van het onderzoek.
- De gebruikte hoeveelheid data moet representatief zijn voor het midden en kleinbedrijf, zodat resultaten direct de echte werkomgeving van Efficiency-Online representeert. Dit

dwingt ook schaalbaarheid af. Naast de benodigde processorcapaciteit zal er ook gekeken moeten worden naar de benodigde schijfruimte voor de datasets.
- De consistentie van de data dient te allen tijde gewaarborgd te blijven.
- De transformatie mag initieel lang duren maar moet incrementeel competitief zijn met standaard web applicaties en met de normale *OLAP* techniek zodat het bijhouden van de data eenvoudig is en alleen het opzetten van een nieuwe analyse meer werk vereist.
- Het systeem zal robuust moeten zijn zodat er geen performance problemen optreden bij zwaar gebruik en zich geen onverwachte problemen voordoen. Bij bijvoorbeeld de transformatie van de data, mag de server niet neergaan.
  De data integriteit wordt hierdoor ook deels gegarandeerd.

Na het uitvoeren van de opdracht zullen de volgende onderdelen opgeleverd worden:
- Een proof of concept van het *WOLAP* tool die niet triviale berekeningen mogelijk maakt
- Een profilering van een groep gebruikers
- Een specificatie van het gehele systeem, waar dit document een basis voor is
- Een gedetailleerd en wetenschappelijk onderbouwt ontwerp
- Het systeem zelf, inclusief broncode, documentatie en eventueel een *API, A*pplication *P*rogrammers *I*nterface.
- Handleidingen voor het opzetten van een profilering en het gebruik van het *WOLAP* tool.

2.3 Behoefte

De behoefte van bedrijven om gecompliceerdere analyses dan bijvoorbeeld simpele webstatistieken online uit te voeren, is de afgelopen jaren enorm gestegen. Bedrijven als Microsoft en Oracle zijn momenteel bezig om de huidige intranet oplossingen om te zetten naar web oplossingen in respectievelijk .NET en de Oracle omgeving. Vaak zijn deze oplossingen te zwaar en financieel te onaantrekkelijk voor het midden- en kleinbedrijf of kleinere afdelingen in grote bedrijven. Efficiency-Online probeert aan deze behoefte te voldoen maar heeft nog geen tool om daadwerkelijk competitief te kunnen zijn.

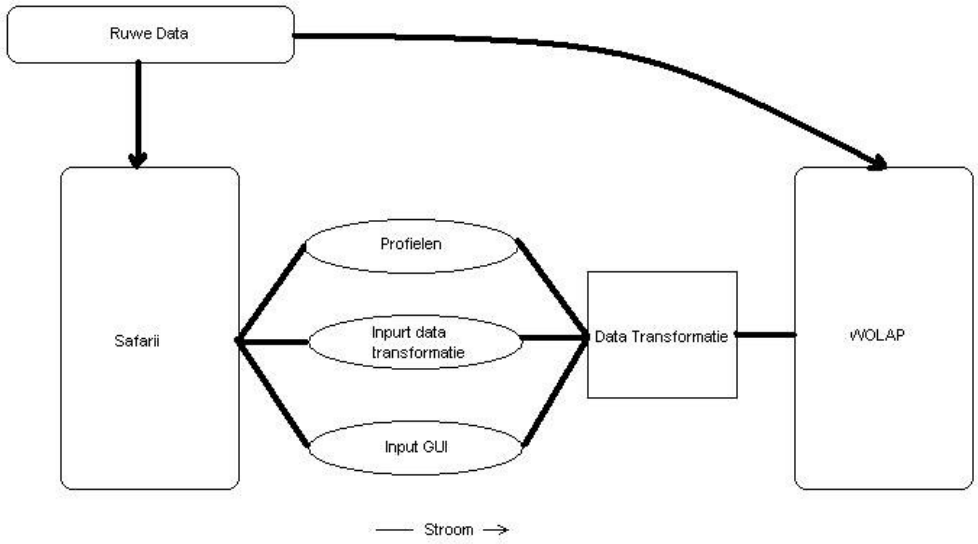
2.4 Inzicht

De uitdaging ligt wat mij betreft geheel in het extra element profilering en de datatransformatie. Door deze methodes met elkaar te combineren probeer je een beter product op te leveren.
Men kan het zien als vals spelen. De eis van alle informatie moet opvraagbaar zijn, zwak je af, door alleen de informatie die past bij het profiel van een gebruiker snel aan te bieden. Een gebruiker merkt hier idealiter niets van, waardoor het doel wordt bereikt.
Tevens wordt men gedwongen bij het maken en onderzoeken van datatransformaties gedwongen om het probleemgebied tot in detail te doorgronden. De kennis en inzichten die tijdens deze fase worden opgedaan zijn zeer waardevol voor het ontwerp en implementatie van het *WOLAP* tool.

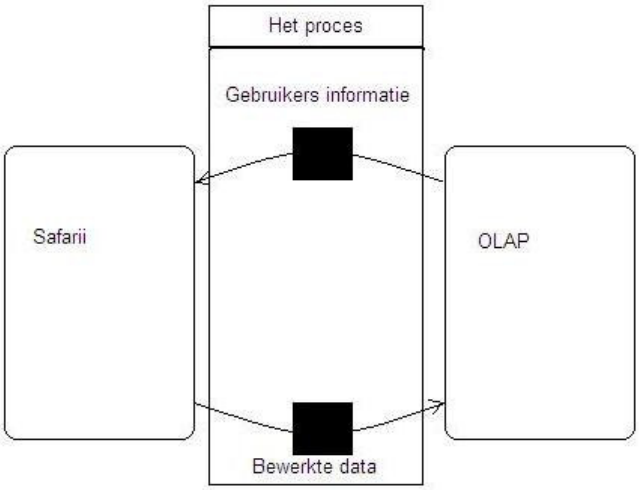
2.5 Afbakening

Het gehele traject wat hier wordt beschreven is te uitgebreid om als geheel te implementeren. De focus ligt dan ook op de datatransformatie van de output van Safarii naar de input van een *WOLAP* applicatie (zie figuur 2.5). Het dataminen en de *WOLAP* applicatie kunnen als bekend worden verondersteld. Dit met de bedoeling de randvoorwaarden voor de transformatie al vervuld te hebben. Mocht de transformatie voorspoediger verlopen dan is er altijd de mogelijkheid om daarna meer de focus te leggen op het *WOLAP* tool en het dataminen. Dit wil natuurlijk niet zeggen dat er niet gekeken wordt naar de functionaliteit van beide methoden; dit zal zeer belangrijk zijn om in het achterhoofd te houden. Het belangrijkste is dat er geen tijd verloren gaat met het ontwerpen of specificeren van methoden die niet van belang zijn voor de datatransformatie.

Uit bovenstaande striktere afbakening kan men afleiden dat het onderzoek als geslaagd kan worden gezien als er een goed ontwerp opgeleverd wordt met koppelingen naar zowel *WOLAP* als Safarii. Tevens dient er een zogenaamd proof-of-concept opgeleverd te worden zodat te zien is of de transformatie afdoende is en om ervoor te zorgen dat implementatie aspecten meegenomen worden in het ontwerp.



Figuur 2.2 *Schematische weergave van het idee en de opdracht*



Figuur 2.5 *Het proces*

3. Proces

In dit hoofdstuk zal het proces worden besproken dat gebruikt zal gaan worden voor dit project. Dit proces is een algemene richtlijn aangezien er momenteel nog geen uitspraken gedaan kunnen worden over de problemen die met de verschillende componenten gepaard zullen gaan. Vandaar dat het proces abstract over zal kunnen komen. We willen een algemeen proces dat door veel organisaties gebruikt wordt gaan doorlopen.

3.1 Oriëntatie

Allereerst zal de oriëntatie fase ingaan. In deze fase zal worden begonnen met de vergaring van één of meerdere representatieve datasets en met de initialisatie van de Safarii software. Nadat deze globale omgeving ingericht is zal er worden begonnen met een onderzoek naar de te gebruiken profileringmethode en datatransformatie. Hierna zal de eerste simpele implementatie gaan plaatsvinden van beide componenten. In deze fase zullen veel methodes uitgeprobeerd worden, men kan dit zien als een empirisch onderzoek om er achter te komen wat de beste methoden zouden kunnen zijn om een start te kunnen maken met het ontwerp. Het idee hierachter is om eerst te weten wat je wilt ontwerpen om het daarna goed te doen.

3.2 Specificatie

Na de eerste globale verkenning van het probleem domein zal in deze fase precies gespecificeerd worden wat er precies gedaan gaat worden en waar de problemen zullen liggen. In deze fase zal duidelijk zijn wat we uiteindelijk in het WOLAP systeem kunnen verwachten aan functionaliteit en zullen de puntjes op de i dienen te worden gezet op het gebied van de profilering en datatransformatie.

3.3 Ontwerp

In tegenstelling tot het wat zal in deze fase de nadruk liggen op het hoe. Hier zal wetenschappelijk worden vastgelegd hoe het systeem gebouwd zal gaan worden. Hier wordt rekening gehouden met de vorige twee fases. Deze dienen namelijk als input voor de te kiezen weg. Naar aanleiding van het empirisch onderzoek uit de oriëntatie fase kan impliciet een deel van het ontwerp worden afgeleidt. Vandaar dat we hier zullen kiezen voor het meer in detail ontwerpen van het *WOLAP* tool.
We willen gebruik maken van een lichte vorm van ontwerpen met uitzondering van de datatransformatie. Deze dient uitvoerig onderzocht te zijn in de vorige fases en gedetailleerd ontworpen te worden in deze fase.

3.4 Implementatie

In deze fase zal het systeem daadwerkelijk gebouwd worden aan de hand van de specificatie en het ontwerp. Ook zal hier de tweede testfase plaatsvinden. Het volledige proces zal hier uitgewerkt worden en het proof of concept zal hier opgeleverd moeten gaan worden. Men kan ervoor kiezen om hierna nog een fase oplevering te doen, maar gezien de interne structuur van de opdracht zal dit deze fase plaatsvinden. Dit houdt in dat de documentatie in deze fase geschreven zal gaan worden.

3.5 Scriptie

Het afsluiten van deze (stage) opdracht zal gebeuren door het schrijven van een scriptie en het houden van een presentatie. In deze scriptie beschrijf ik precies wat ik tijdens mij stage allemaal gedaan heb, welke keuzes ik gemaakt heb en waarom, welke problemen ik ben tegengekomen en hoe ik deze opgelost heb.
Ik ga proberen de scriptie zoveel mogelijk samen te stellen tijdens de eerder genoemde fases zodat er duidelijk aan bod zal komen welke problemen wanneer speelden. Dit zal hopelijk voorkomen dat er achteraf onduidelijkheid bestaat over de volgorde en detail van onderdelen die beschreven worden.

4. Planning

In dit hoofdstuk zal een globale planning worden besproken. Ik zal me beperken tot een globale planning per fase. Een gedetailleerde planning is mijns inziens niet te maken aangezien er veel afhankelijkheden zijn waar ik geen invloed op heb. De globale planning zal ik proberen strikt te volgen, aangezien dit toch een beeld geeft van de geschatte tijd per onderdeel.

Weeknummer van:	**Weeknummer tot:**	**Fase:**
37	40	Plan van Aanpak
37	44	Oriëntatie
45	47	Specificatie
48	52	Ontwerp
53	5	Implementatie
5	10	Scriptie

Uit hoofdstuk 2 en 3 blijkt dat er een cyclisch verloop zal zijn in orientatie specificatie en ontwerp die iedere ronde een proof-of-concept opleveren. Dan wel via een document dan wel via een implementatie. Dit zorgt voor de waarborging van het opleveren van een proof-of-concept.

5. Literatuurstudie

Tijdens de plan van aanpak / oriëntatie fase heb ik al beperkt onderzoek gedaan naar relevante literatuur en systemen en zal dit verder uitbouwen tijdens de oriëntatie / specificatie fase. Hoewel een groot deel van het onderzoek naar literatuur en mogelijkheden pas tijdens het uitvoeren van het project plaatsvindt, vormen onderstaande punten een goede basis om de oriëntatie / specificatie fase mee te beginnen. De belangrijke onderdelen van dit onderzoek worden hieronder beschreven. In paragraaf 5.6 zal ik een voorbeeld schetsen om de besproken literatuur in perspectief te plaatsen.

5.1 Standaard OLAP transformaties

Aangezien het systeem gebruik gaat maken van *OLAP* en er een standaard transformatie wordt gebruikt voor deze systemen heb ik deze transformaties met de te gebruiken software (PHP samen met MySQL) geïmplementeerd op een aantal verschillende manieren om te kijken wat de effecten hiervan zijn. Deze standaard transformatie heet Starscheme[3]. Hieronder zal stapsgewijs dit proces weergegeven worden:

- Maak heuristieken om te bepalen welke kolommen relationeel worden en welke niet
- Onderzoek of deze heuristieken voldoen aan de te gebruiken database, pas ze zo nodig aan
- Onderzoek welke kolommen een zoekindex krijgen op basis van de te gebruiken database
- Maak het lege datamodel volgens de dataset
- Vul dit datamodel met de gegevens van de dataset. Hierbij mag geen informatie verloren gaan uit de data

Er zijn een aantal basisideeën voor dit algoritme:

- Hoe smaller een tabel hoe beter, dus ga je al snel vervallen in een relationele structuur
- Rekenen met integers gaat sneller dan met andere datatypen
- Iedere database heeft zijn eigen optimalisaties en methoden om deze te gebruiken

Het zal duidelijk zijn dat dit alleen aangeeft wat er als basis moet zijn om überhaupt te kunnen beginnen aan de ontwerpfase en implementatie fase. Deze ideeën geven weer waar bij de profilering van gebruikers op gelet dient te worden.

Bij het opstellen van de heuristieken kan men al denken aan transformatie. Ik heb deze heuristieken strikt simpel gehouden om te voorkomen dat er bij de keuze van transformaties te snel gekozen zal worden voor een heuristische aanpak.

5.2 Standaard OLAP functionaliteit

Er kan gesteld worden dat via een *OLAP* tool altijd van iedere view naar iedere andere view[1] gesprongen kan worden en wel zodanig dat de eerdere informatie behouden blijft of meegenomen wordt in een volgende view. Deze standaard functionaliteit heet Drill-Down. Wil men bij een

[1] Een view is een doorsnijding van de data. Bijvoorbeeld bij web statistieken het aantal (operator) bezoekers (metriek) per jaar (x-as) per continent (y-as).

bepaalde view alle overige relevante informatie zien, moet dit ook een mogelijkheid zijn. Je keert terug naar de ruwe tabel om heel direct te zijn. Deze functionaliteit heet Drill-Through.
Beide functionaliteiten dienen beschikbaar te zijn in het proof of concept mocht de datatransformatie het gewenste resultaat behalen in de gewenste tijd.

5.3 Algemeen

Inmiddels heb ik al onderzoek gedaan naar implementatie technieken en optimalisaties in de MySQL database. Deze technieken zijn talrijk en zijn los alleen maar gereedschap om het ontwerp te kunnen implementeren, dus deze zullen niet los besproken worden. Ik zal me beperken tot een klein aantal concepten:

- Relationele database: dit zal de standaard opslagmethode zijn
- Serialization: dit is het representeren van objecten als string en het opslaan van deze string; dit mechanisme wordt door PHP aangeboden. Dit zal tijdens het maken van WOLAP intern gebruikt gaan worden om gegenereerde tabellen in geheugen op te slaan.
- Queues: tijdens algemene gesprekken met medestudenten is dit een mogelijke manier gebleken om het incrementeel opslaan van data in het datamodel te bespoedigen.

5.4 Gebruikers

Een groot probleem kan het vergaren van gebruikersdata worden. Hier hebben we de mogelijkheid onderzocht om deze random te generen met een wisselende frequentie per beoogd profiel. Hier zal tijdens de orientatie fase nog dieper naar gekeken moeten worden. Tevens is er een mogelijkheid gebruik te maken van ‘real time’ data, maar hiervoor dienen we contact op te nemen met enkele van de klanten van Efficieny-Online. Deze laatste optie geniet natuurlijk de voorkeur aangezien je dan de mogelijkheid hebt om uitspraken te doen over de kwaliteit van het systeem in de ‘echte’ wereld.

5.5 Transformatie

Het belangrijkste onderdeel van mijn stage is de transformatie van de output van het datamining tool naar goede input voor een *WOLAP* tool. Omdat ik op dit terrein nog redelijk onbekend ben heeft Patrick van Bommel me een handreiking gegeven in de vorm van een transformatie methode die erg abstract is en zeer generiek te gebruiken is. De transformatie methode is een graaf transformatie. Dit wil zeggen dat er in een graaf S, reducties, de transformatie regels, zonodig herhaaldelijk worden toegepast om bij een gewenste graaf E uit te komen. In 5.5.1 en 5.5.2 definieer ik een basis elementen.

5.5.1 Grafen

Een definitie van een graaf[2].

Een graaf G bestaat uit 2 verzamelingen; V en E.

Voor V geldt:
- V is niet leeg
- Alle elementen uit V noemen we punten

Voor E geldt:
- Elk element uit E is een ongeordend paar van unieke elementen uit V
- E = V * V
- Alle elementen uit E noemen we lijnen

5.5.2 Graaf transformaties

Bij graaf transformaties wordt vaak gebruik gemaakt van een gelabelde, gerichte graaf[3].

Een gelabelde gerichte graaf G: G = (V, E, source, target, label)

Voor V geldt:
- V is niet leeg
- Alle elementen uit V noemen we punten

Voor E geldt:
- Elk element uit E is een ongeordend paar van unieke elementen uit V
- E = V * V
- Alle elementen uit E noemen we lijnen

Voor source geldt:
- Dit is het element dat bij iedere lijn een start punt aangeeft

Voor target geldt:
- Dit is het element dat bij iedere lijn het eind punt aangeeft

Voor label geldt:
- Dit is het element dat bij ieder punt een label toewijst

Een graaf transformatie komt overeen met het toepassen van een regel of regels op een graaf en herhaling van dit proces totdat de regel of regels niet verder toegepast kunnen worden op de desbetreffende graaf.

Een regel[3]:

Een regel R: R = (L, R, K, glue, emb, appl)

Voor L, R geldt:
- Een linker en rechter graaf

Voor K geldt:
- We noemen dit de interface graaf
- K is een deelgraaf van L

Voor glue geldt:
- Een voorkomen van K in R, relateert K met R

Voor emb geldt:
- Verankeringsrelatie, relatie tussen punten uit L & punten uit R

Voor appl geldt:
- Toepassingscondities voor de regel

5.6 Voorbeeld

Laten we een simpele case beschouwen. Laten we uitspraken willen doen over de volgende data.

Ruwe data tabel:

Id	Continent	Land	Jaar	Maand	Website
1	Europa	Nederland	2004	10	www.google.nl
2	Azië	Japan	2004	5	www.kun.nl
…	…	…	…	…	…

We zien hier 6 dimensies (= aantal kolommen) die tegen elkaar uitgezet kunnen worden en die alle 6 als metriek kunnen dienen. Er zijn dus 6^3 = 216 mogelijke views te maken. Dit geeft aan hoeveel informatie er uit deze simpele dataset te halen is, wat niet wil zeggen dat dit altijd nuttige informatie zal zijn.
Er zitten twee mogelijke Drill-Down functionaliteiten in de dataset: van jaar -> maand en van Continent -> Land.

Mogelijke profileringen:
Profilering 1:
- Groep geïnteresseerd in land, jaar en aantal hits
- Groep geïnteresseerd in alles wat er te halen valt

Profilering 2:
- Groep geïnteresseerd in de helft van alle dimensies
- Groep geïnteresseerd in alles wat er te halen valt

…

Uit deze bewust gekozen profilering kan men de conclusie trekken dat er nog veel in te vullen is aan specifieke details over welke profilering nu het beste past bij *OLAP* en de datatransformatie.

Nu zijn aan de voorwaarden voldaan. Er is een ruwe tabel er is een profilering en we weten welke doorsnijdingen er mogelijk zijn.

6. Onderzoeksvraag

In dit hoofdstuk zal de onderzoeksvraag worden besproken. Zoals in dit document is gebleken staan een aantal onderwerpen centraal:

- Profilering
- Data transformatie
- Data doorsnijding door middel van WOLAP

Om deze verscheidenheid om te vormen in een onderzoeksvraag hebben we ervoor gekozen om het basisidee van de opdracht als leidraad te gebruiken.
Hiermee wordt de centrale onderzoeksvraag:

"Is het mogelijk om een algoritme te ontwikkelen dat via datatransformaties en datamining, een niet triviale gebruikersprofilering oplevert, dat als invoer kan dienen om standaard *OLAP* functionaliteit efficiënter en sneller aan te kunnen bieden."

Uit de onderzoeksvraag kan worden afgeleidt dat het daadwerkelijk gaat om het concept en zal een proof of concept van een *OLAP* applicatie voldoende zijn om aan te kunnen tonen of de doelstelling bereikt is of kan worden bereikt.

7. Literatuurlijst

De literatuurlijst die is gebruikt voor het maken van dit plan van aanpak.

1. **Graph Transformation for Specifcation and Programming**
   Marc Andries, Gregor Engels, Annegret Habel, Bertholf Hoffman, Hans-Jörg Kreowski, Sabine Kuske, Detlef Plump, Andy Schürr, Garbiele Taentzer

2. **Discrete Mathematics with Graph Theory**
   Edgar G. Goodaire, Michael M. Parmente

3. **http://explanation-guide.info/meaning/OLAP.html**
   Wikipedia onder de voorwaarden van GNU FDL