DATAMODELLERING TOEPASSEN DATA ANALYTICS

Inleiding

In dit whitepaper wordt een toepassingsgebied beschreven voor datamodellering. Een toepassing is een werkveld op het vlak van architectuur of modellering waarbij een aantal data modelleervormen met elkaar gecombineerd worden.

Deze specifieke modelleervormen zijn beschreven in een serie whitepapers. In de whitepapers over toepassingsgebieden gaan we in hoe de verschillende modelleervormen met elkaar gecombineerd worden ter ondersteuning van dit toepassingsgebied

Deze combinatie maakt het vervolgens mogelijk om op adequate wijze een model te communiceren voor dit toepassingsgebied. In een aantal gevallen wordt alleen documentatie geproduceerd, in andere situaties kunnen ook andere zaken geproduceerd worden zoals source code of templates etc.

Doel

Data Analytics kent veel verschijningsvormen, denk aan Business Intelligence, Data Science, Analytics of Machine Learning. Data Analytics is een werkveld dat momenteel volop in beweging is, er ontstaan nog steeds nieuwe vormen van analysemethoden en -technieken. Daarnaast vinden op technologisch vlak veel ontwikkelingen plaats. Denk aan Big Data platformen, NoSQL databases en vormen van analytics systemen.

Datamodellering speelt echter in alle vormen van analytics een belangrijke rol. Met name het objectmodel dat gebruikt wordt als onderliggend model voor de analyse is een vorm van data modellering. Soms is dit zeer expliciet zoals in een Data Ware House (DWH), bij andere toepassingen is dit meer impliciet zoals in tools voor data scientists.

Wordt data geproduceerd, verwerkt, opgeslagen en getransporteerd dan kan deze data geanalyseerd worden. In een aantal gevallen is de verwerking specifiek voor analytics (DWH). Andere situaties kenmerken zich dat de data eerst verzameld en opgeslagen is voor andere toepassingen en vindt later analyse plaats. Echter in al deze situaties is een datamodel aanwezig.

In dit whitepaper behandelen we de Data Analytics in meer algemene zin en niet gericht op een specifiek vakgebied. Sommige notatiewijzen zijn meer expliciet voor één modelleervorm, anderen zijn meer algemeen.

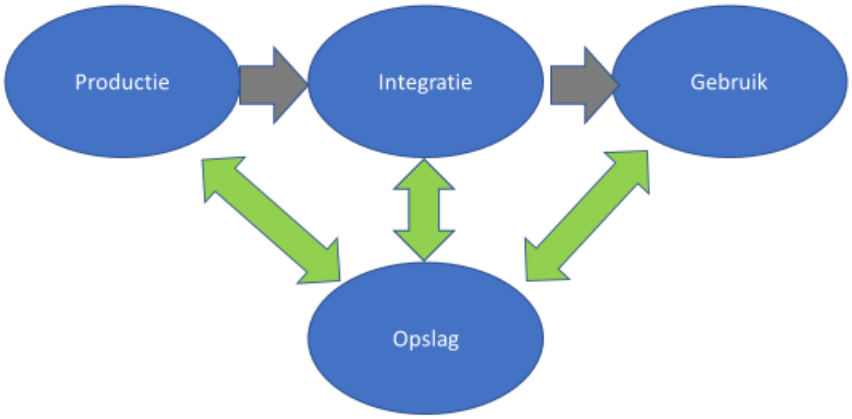
Context

Data analytics is binnen sommige organisaties goed ontwikkeld. Hier werken verschillende disciplines zoals data engineers, vakinhoudelijk experts en data scientists nauw met elkaar samen om te zoeken naar patronen in dataverzamelingen waar de organisatie zich mee kan onderscheiden ten opzichte van concurrenten.

In andere organisaties is data analytics eenvoudiger van opzet en minder ontwikkeld en bestaat analytics uit een aantal eenvoudige rapportages binnen de productionele informatiesystemen. In alle situaties is

een model relevant. Er wordt hierbij vaak gesproken over “schema on read en schema on write”. Een belangrijk concept in nieuwe toepassingen waarbij schema betrekking heeft op het datamodel dat voor de analyse ingezet wordt.

Voor datamodellering en analytics verwijs ik meestal naar een eenvoudig model gebaseerd op de informatie theorie van Shannon.



Dit model geeft aan dat er data geproduceerd wordt en gebruikt en hoe groter de afstand tussen productie en gebruik hoe meer integratie en opslag plaats zal vinden. Voor data analytics kan voor iedere stap in dit model een data model opgesteld worden. In een volwassen organisatie is dit essentieel om op adequate wijze de juiste bronnen voor data analyse te kiezen en de data in deze bronnen op efficiënte wijze te converteren naar een model geschikt voor analyse.

DOELEN VAN DATA ANALYSE

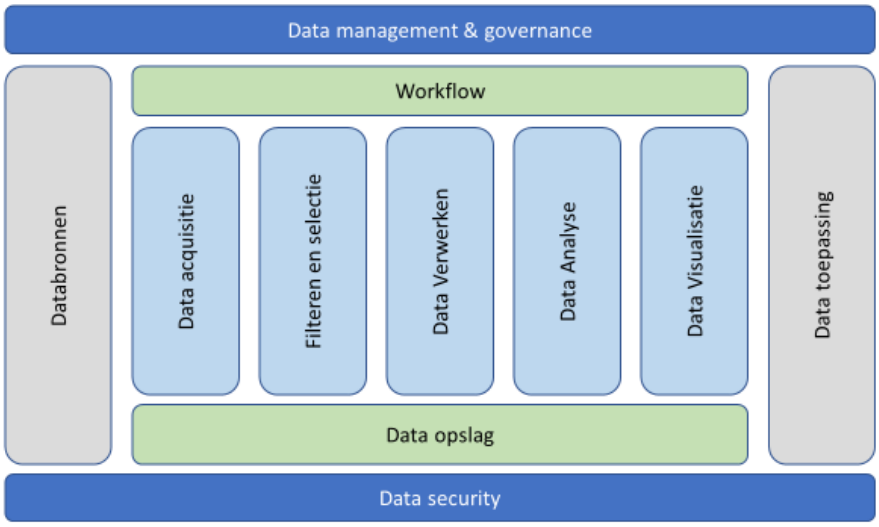
Data analyse heeft vele doelen, onderstaande opsomming geeft een niet uitputtend overzicht:

- **Optimalisatie van werkprocessen**, als analyse vormen worden ingezet kan een werkproces anders vormgegeven worden. De handmatige- of semi geautomatiseerde activiteiten kunnen (deels) vervangen worden door algoritmes op dataverzamelingen van binnen en buiten de eigen databronnen.
- **Signalering,** op een voldoende vroeg moment signaleren van abnormaliteiten in data(stromen) kunnen een organisatie ondersteunen bij het op juiste tijd nemen van beslissingen of ondernemen van corrigerende acties.
- **Onderzoek en wetenschap,** in de wetenschap is het doen van data analyse niet meer weg te denken. Veel analyse technieken zijn ontwikkeld in het werkveld van onderzoek en wetenschap.

- **Beslissingen nemen,** data wordt omgezet naar informatie en dat wordt vervolgens omgezet naar kennis. Met de juiste kennis kunnen onderbouwde beslissingen genomen worden.

Analytics en data modellering

In een vorige paragraaf hebben we reeds de levensloop getoond van data. Hieronder geven we een meer uitgewerkt raamwerk dat ingezet kan worden in Big Data toepassingen en data analytics. Ook voor DWH en BI omgevingen is dit raamwerk bruikbaar.

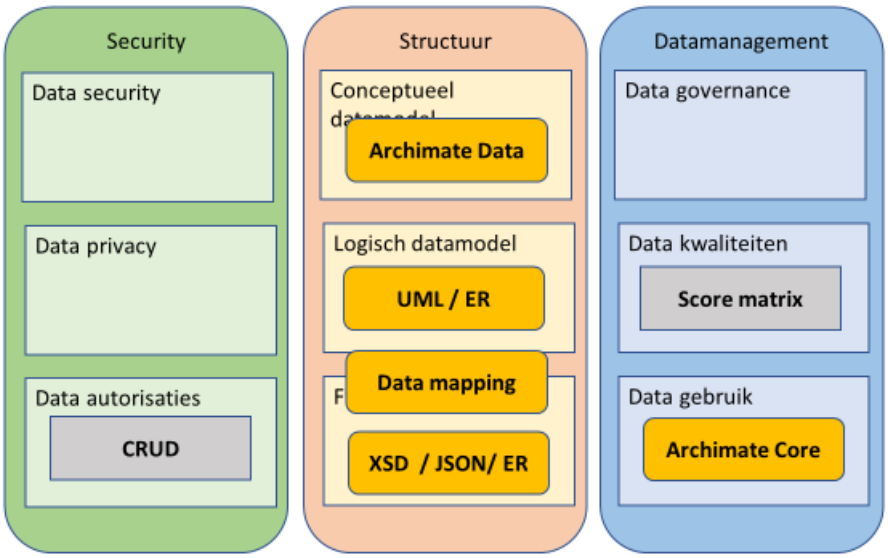


1

Voor iedere stap kunnen een aantal notatiewijzen ingezet worden. Een aantal zijn reeds behandeld in andere whitepapers zoals data management, security en opslag. Dit model kan ingezet worden om de deelactiviteiten of functionaliteiten te plotten op het model en dit vervolgens uitwerken tot een transformatiemodel wat de modelveranderingen in iedere fase laat zien. Dit eventueel in combinatie met de componenten die daarvoor ingezet gaan worden.

Notatiewijzen

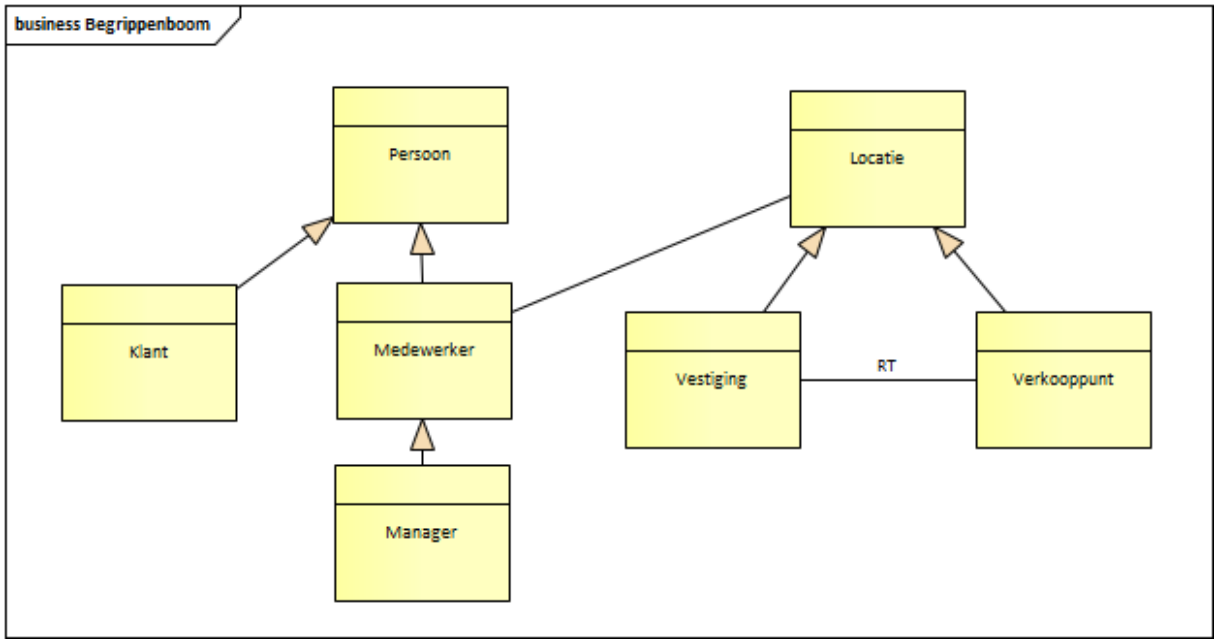
Voor data modellering binnen data analytics zijn een aantal notatiewijzen relevant. Een aantal is essentieel, en een aantal is ondersteunend. Onderstaande afbeelding geeft een beeld van de notatiewijzen die vervolgens kort worden toegelicht.



CONCEPTUEEL DATAMODEL

Het conceptueel datamodel is voor data anlytics een essentieel onderdeel dat zorgt voor de beschrijving welke data entiteiten cq data sets relevant zijn binnen de organisatie en hoe deze zich verhouden tot elkaar en tot het gebruik in de organisatie.

Onderstaande afbeelding geeft een beeld van een eenvoudig conceptueel datamodel uitgewerkt binnen de ArchiMate notatie. Meer informatie over de notatiewijze is te vinden via: http://assistent.interactory.nl/cmsForm.aspx?formid=50027&webcontentid=248

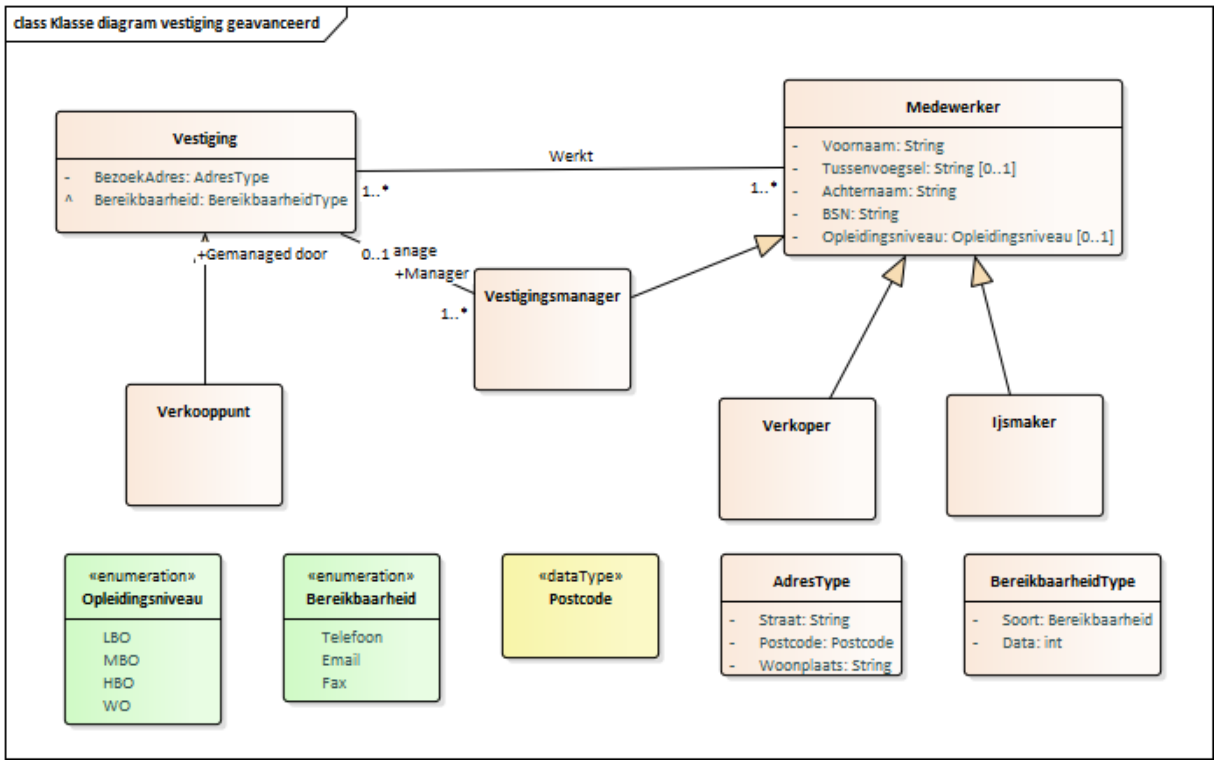


LOGISCH DATAMODEL

Het logisch model is in data analytics voornamelijk van belang bij het beschrijven van de structuur zonder rekening te houden met de fysieke implementatie. Dit model is de basis van het analyse model en dient daarom voldoende detail te bevatten om een analyse te kunnen uitvoeren.

Onderstaande afbeelding geeft een voorbeeld van deze notatiewijze. Er zijn echter meerdere vormen van notatie mogelijk, zoals het ER Diagram. De onderstaande link geeft een beeld van het UML klasse diagram.

http://assistent.interactory.nl/cmsForm.aspx?formid=50027&webcontentid=259



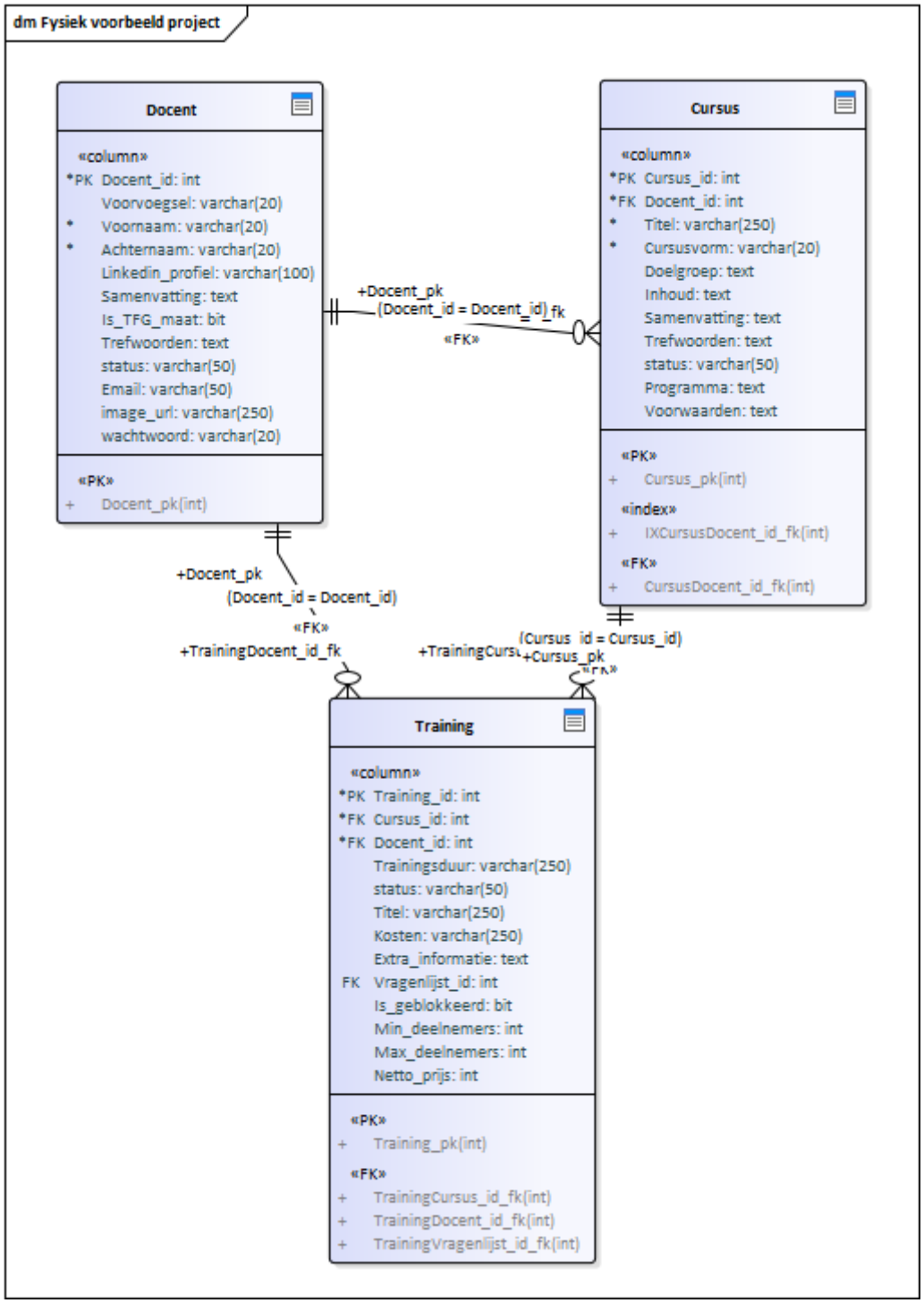
FYSIEKE MODELLERING

De fysieke modellering wordt voornamelijk gebruikt om de structuur van de databronnen te beschrijven. Dit wordt gedaan met een notatiewijze die aansluit bij de structuur van de data in de databronnen. Denk bijvoorbeeld aan ER diagrammen voor relationele databases, XSD structuren voor NoSQL databases of het gebruik van data integratie systemen als bron voor een data analyse.

Naast de (semi) gestructureerde datasets wordt binnen data analyse steeds vaker gebruik gemaakt van datasets met weinig structuur. Denk bijvoorbeeld aan de inzet van social media voor het doen van analyse. Een aantal social media platformen ontsluiten met name laag gestructureerde data. Modellering van deze laag gestructureerde data is een uitdaging en wordt veelal op basis van andere dimensies gedaan. Denk bijvoorbeeld aan verschillende tellingen van woorden.

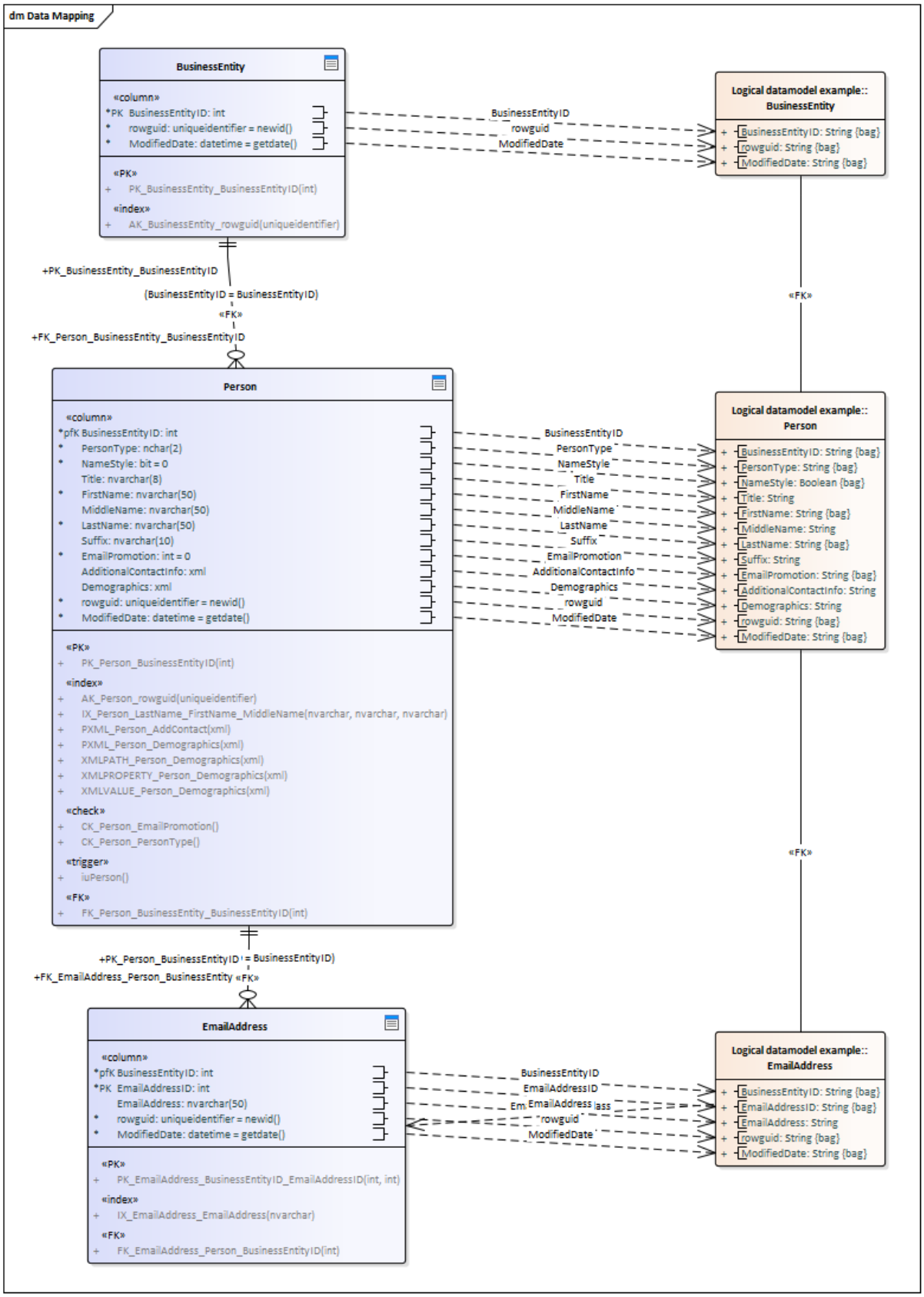
Onderstaande afbeelding geeft een beeld van een gestructureerd datamodel uit een relationele database. Meer informatie is te vinden via de volgende whitepapers.

- http://assistent.interactory.nl/cmsForm.aspx?formid=50027&webcontentid=253
- http://assistent.interactory.nl/cmsForm.aspx?formid=50027&webcontentid=261

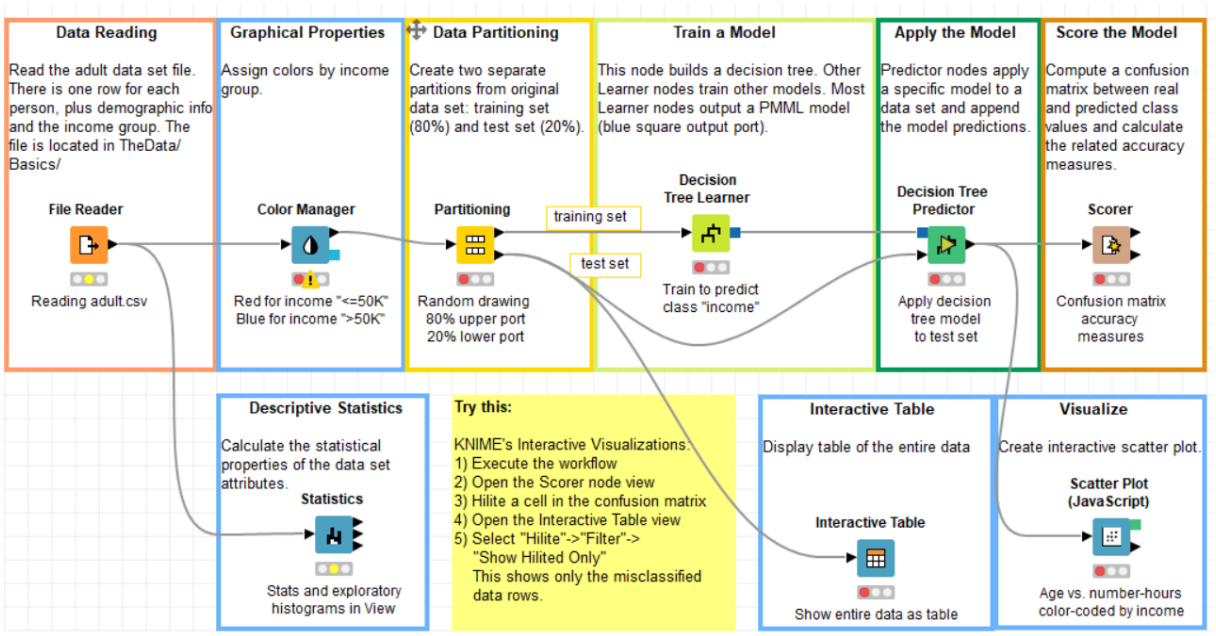


Naast de structuur van de data bronnen en het (logische) model voor de analytics is er een derde vorm van modelleren van groot belang. Dat is de modellering van de data in beweging en de uitwerking van de transformaties en bewerkingen die op de data plaatsvinden.

Hiervoor wordt veelal een vorm van data mapping gebruikt. Hierbij wordt een aangepaste vorm van ER of UML toegepast, zoals te zien is in onderstaande afbeelding.



Daarnaast zie je dat tools voor data analyse en – integratie vaak een eigen notatiewijze hebben die gebaseerd zijn op een vorm van data mapping. Een aardig voorbeeld in de afbeelding hieronder vanuit de Open Source tool Knime:



Hierbij zie je feitelijk alle stappen zoals beschreven in het raamwerk terug. Van de bronnen via de transformaties naar een visualisatie en desgewenst de opslag van de gegevens in tussen- en eind data opslag. Meer informatie over data mappings is hier te vinden:

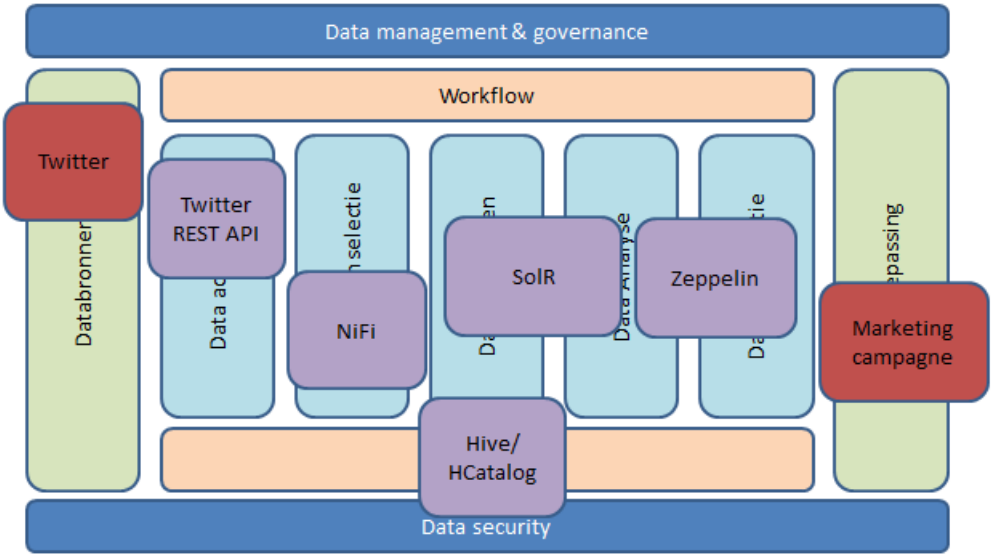
http://assistent.interactory.nl/cmsForm.aspx?formid=50027&webcontentid=252

DATA GEBRUIK

Data analytics is een specifieke vorm van data gebruik. Het inzichtelijk maken welke analyses gebruik maken van welke datasets kan daarom toegevoegde waarde bieden. Het geeft een beeld welke dataset vaak ingezet wordt voor analyse. Dat onderstreept de relevantie van de data maar mogelijk ook van de kwaliteiten van de data in deze set.

Data gebruik kan ingezet worden het modelleren van de datasets die ingezet worden maar ook voor de transformaties die gedaan worden op een dataset tussen bron en toepassing. Hiervoor kan de ArchiMate core goed ingezet worden. Bijvoorbeeld de applicatie functies en de componenten. Veelal zal dit geplot worden op het raamwerk zoals beschreven eerder in dit hoofdstuk. Meer informatie over de notatiewijze is te vinden via:

http://assistent.interactory.nl/cmsForm.aspx?formid=50027&webcontentid=246



Data autorisaties

Data autorisaties zijn optioneel voor data analytics. Dit is afhankelijk van de context van de organisatie maar ook van het soort datasets dat ingezet wordt voor de analyse. Zijn dit datasets met een hoge mate van vertrouwelijkheid of met privacy gevoelige data dan is het wenselijk om in kaart te brengen wie toegang heeft tot deze datasets (en de eventuele tussenproducten).

Hiervoor is een CRUD matrix een goed hulpmiddel. Een CRUD matrix geeft aan per dataset welke bedrijfsrol, lees, creer, muteer en verwijder rechten heeft. Meer informatie over de CRUD matrix is te vinden in whitepaper: http://assistent.interactory.nl/cmsForm.aspx?formid=50027&webcontentid=250

Onderstaande afbeelding is een voorbeeld van een eenvoudige CRUD matrix:

Target / Source	Manager	Medewerker	Salarisadministratie
Manager	CR		CR
Medewerker	CRUD	CRUD	
Uur	CRUD	CRUD	R
Vestiging	CRUD	CRUD	

DATA QUALITY

Voor het modelleren van de kwaliteit van een dataset binnen een analyse situatie is in eerste instantie een overzicht van relevante data kwaliteiten van belang. Deze lijst kan men zelf opstellen, echter er zijn reeds een aantal standaard kwaliteitsindelingen aanwezig. Het model van DaMa is het meest compleet en wordt daarom veel ingezet.

Is er een lijst van kwaliteiten gedefinieerd dan kan deze gerelateerd worden aan de verschillende datasets binnen de analyse. Hiervoor zijn meerdere mogelijkheden, de meest eenvoudige, maar toch effectieve, aanpak is een scorematrix. Hierbij geef je in de matrix aan hoe een dataset scoort op een bepaalde kwaliteit. Datasets met een hoge score in kwaliteiten kunnen goed ingezet worden bij de selectie van datasets in de aquisitie fase binnen het analyse raamwerk.

De score matrix een goed hulpmiddel. Score kan ingevuld worden met een getal tussen 0 en 10 of een ordinale indeling zoals Laag – Midden Hoog. Detailinformatie over de modelleervorm score matrix is te vinden via: http://assistent.interactory.nl/cmsForm.aspx?formid=50027&webcontentid=256

Target / Source	Accuraatheid	Actualiteit	Compleetheid	Consistentie	Precisie	Privacy	Redelijkheid	Referentiele integriteit	Tijdigheid	Uniekheid	Validiteit
Cursus	8	8	8	8	NaN		NaN	NaN	8	9	6
Docent	8	8	8	8	NaN	6	NaN	NaN	8	6	6
Training	8	8	8	8	NaN		NaN	NaN	8	9	6

Kenmerken

Data Analytics worden door steeds meer organisaties ingezet als middel om nieuwe inzichten te verkrijgen uit data en deze techniek komt daarom bij steeds meer organisaties hoger op de prioriteitenlijst te staan. Daarnaast neemt de hoeveelheid data die door organisaties verwerkt worden verder toe wat aanvullende mogelijkheden biedt voor analytics.

Big Data en Data Analytics bieden vanuit data modelleringsperspectief een aantal interessante modelleerbehoeften, met name de logische datamodellering en de data mappings zijn hierbij essentieel in combinatie met het conceptuele model. Bij de introductie van data analytics zijn de volgende kenmerken relevant:

- Kijk naar een toepassing of onderzoeksvraag relevant in de organisatie en stel hier een eenvoudig model van op.
- Zoek sponsoren cq ambassadeurs op management niveau voor de hierboven genoemde onderzoeksvraag.
- Selecteer de benodigde databronnen en acquireer de data. Stel ook van de bronnen een datamodel op
- Richt iteratief de oplossing in en betrek in een vroeg stadium de juiste (bedrijfsmatige) stakeholders hierbij.

- Stel de data mappings op of selecteer een tool die deze data mappings kan genereren.
- Lever indien relevant een oplossing op en zorg dat deze ingebed wordt in de ICT organisatie.

Producten

De producten van data analytics vanuit data modelleringsperspectief zijn samengevat:

- Conceptueel datamodel
- Logisch datamodel
- Fysieke modellen van de bronnen
- Data mappings
- Modellen rond datagebruik
- CRUD matrix

Tooling

Zoals reeds genoemd zijn er rond data analytics meerdere producten te vinden, veelal als onderdeel van een Big Data platform of data analytics suite. Er zijn vele specifieke producten zoals bijvoorbeeld RapidMiner, Knime, Cloudera, HortonWorks of Elasticsearch/Kibana.

Als laatste is het inzetten van generieke (enterprise) architectuurtooling te noemen. Een aantal architectuur tools hebben de mogelijkheid om datastructuren en modelleertalen met elkaar te combineren waardoor de (data) modelleerbehoefte voor data analytics grotendeels kan worden afgedekt.

Evaluatie

Data analytics is een nieuw vakgebied dat door steeds organisaties wordt ingezet. Er zijn vele vormen van data analytics beschikbaar zoals BI, DWH, Predictive Analytics of Machine Learning.

Binnen data analytics speelt data modellering een rol. Met name het leggen van verbanden tussen de data entiteiten in de bronnen en het logische model van de analyse is essentieel. In een vroeg stadium nadenken welke modelleervormen relevant zijn, hoe deze aan elkaar verbonden worden en hoe de stakeholders daarbij betrokken zijn ondersteunt de introductie van effectieve analytics.

In dit whitepaper hebben we een combinatie van modelleervormen beschreven die een (minimale) set is van generieke notatiewijzen op basis waarvan data analytics in organisaties gemodelleerd kunnen worden. Voor specifieke toepassingen kunnen specialistische modelleervormen nodig zijn.

Over de auteur

Bert Dingemans is trainer op het vlak van data architectuur, data management en Big Data. Hij heeft een passie voor modelleren, modelleertools en het effectief inzetten van geautomatiseerde hulpmiddelen om modellen effectief in te zetten in de praktijk. Bert is te bereiken via bert@interactory.nl