INLEIDING INFORMATIE- EN DATAMODELLERING

Inleiding

Informatie- en datamodellering is een belangrijk onderdeel van veel werkvelden in de bedrijfskunde en de informatiekunde. Denk bijvoorbeeld aan informatie analyse, data of enterprise architectuur, systeemontwerp, gegevensuitwisseling, database ontwerp en datastroommodellen.

Door deze verschillende verschijningsvormen is er in de loop der jaren een groot aantal modelleerwijzen ontstaan rond informatie- en datamodellering. Elke modelleerwijze heeft eigen specifieke kenmerken waarmee ze enerzijds inzetbaar zijn voor bepaalde groepen stakeholders en anderzijds inzetbaar zijn om een bepaald aspectgebied van data inzichtelijk te maken.

Bij het uitwerken van informatie- of datamodellen kan daarom gezocht worden naar welke modelleervorm het beste aansluit bij de doelgroep cq stakeholders maar ook het duidelijkst de relevante aspecten helder in kaart brengt. Het kiezen van de juiste modelleerwijze kan in deze bijdragen aan het op adequate wijze overdragen van informatie omtrent het domein dat gecommuniceerd wordt aan de stakeholders.

Gezien de vele modelleervormen, aspectgebieden en stakeholders is het vinden van een goede modelleervorm niet altijd eenvoudig. Reden om binnen de architectuur assistent een aantal whitepapers te ontwikkelen waarin de belangrijkste datamodelleervormen worden behandeld op basis van een eenvoudig raamwerk.

Dit whitepaper is een introductie op de serie van datamodelleervormen en beschrijft de kaders en uitgangspunten die we gaan gebruiken bij het beschrijven van de verschillende modelleervormen. Naast deze whitepapers is er een training ontwikkeld waarin de verschillende modelleervormen op praktische wijze worden toegelicht: https://www.itmg.nl/training/masterclass-data-en-informatie-modellering/

Historie

Data is al eeuwenoud, misschien zelfs al vanaf het eerste moment dat we zaken gingen opschrijven. Door de jaren heen is de hoeveelheid data en de complexiteit van datastructuren steeds groter geworden. Door de komst van geautomatiseerde toepassingen is deze ontwikkeling explosief toegenomen.

Sinds enige jaren is het concept Big Data geïntroduceerd. Dit concept geeft aan hoe omvangrijk data tegenwoordig is. Echter naast de hoeveelheid zijn ook de snelheid en de veelvormigheid van data complicerende factoren geworden (3Vs). Via de link https://www.winshuttle.com/big-data-timeline/ krijg je een aardig beeld van de ontwikkelingen rond data en big data door de tijd heen te zien.

Door deze verhoging van de complexiteit is de behoefte aan het opstellen van informatie- en datamodellen meegegroeid. De eerste vormen van data modellering stammen uit de jaren 60. Zo is door Codd in 1969 het relationeel datamodel geïntroduceerd (https://en.wikipedia.org/wiki/Relational_model). Dit model wordt nog steeds toegepast in het database werkveld.

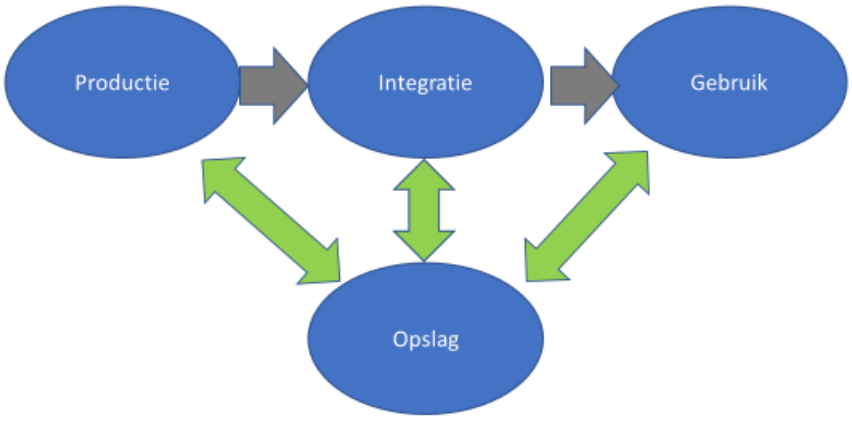
Naast de relationele modellen zijn er door de jaren heen veel nieuwe datamodelleertechnieken ontstaan. Denk aan object georiënteerde modellen zoals UML of flow diagrammen zoals DFD. Zo is er een veelheid aan verschillende informatie- en datamodelleringswijzen ontstaan. In een serie van whitepapers gaan we binnen de architectuur assistent een aantal van deze modelleerwijzen beschrijven op basis van een standaardindeling.

Data levensloop

Naar informatie en data kan op veel manieren gekeken worden. Veel modelleervormen gaan uit van een bepaald gezichtpunt. Echter in deze serie van whitepapers willen we graag raamwerken gebruiken die in elke situatie gebruikt kunnen worden om de verschillende concepten in een modelleerwijze op te plotten.

In de afgelopen jaren ben ik een aantal raamwerken gaan inzetten die het mogelijk maken om verschillende aspecten van data modellering af te beelden. In deze serie artikelen gebruiken we er drie die we in deze en komende paragrafen nader zullen toelichten. Bij het uitwerken van een modelleerwijze gebruiken om op een gestandaardiseerde wijze de modelleervorm te beschrijven.

Het eerste raamwerk is de data levensloop. Deze levensloop toont hoe data in een beperkt aantal stappen verschijnt. Onderstaande afbeelding toont het raamwerk.



Toelichting op het raamwerk:

- **Productie**, data wordt geproduceerd en ontstaat daardoor. Dat kan op vele manieren zijn. Denk hierbij aan personen die gegevens invoeren via formulieren, logs en dergelijke in informatiesystemen of devices zoals smart meters en mobiele telefoons die gegevens produceren.

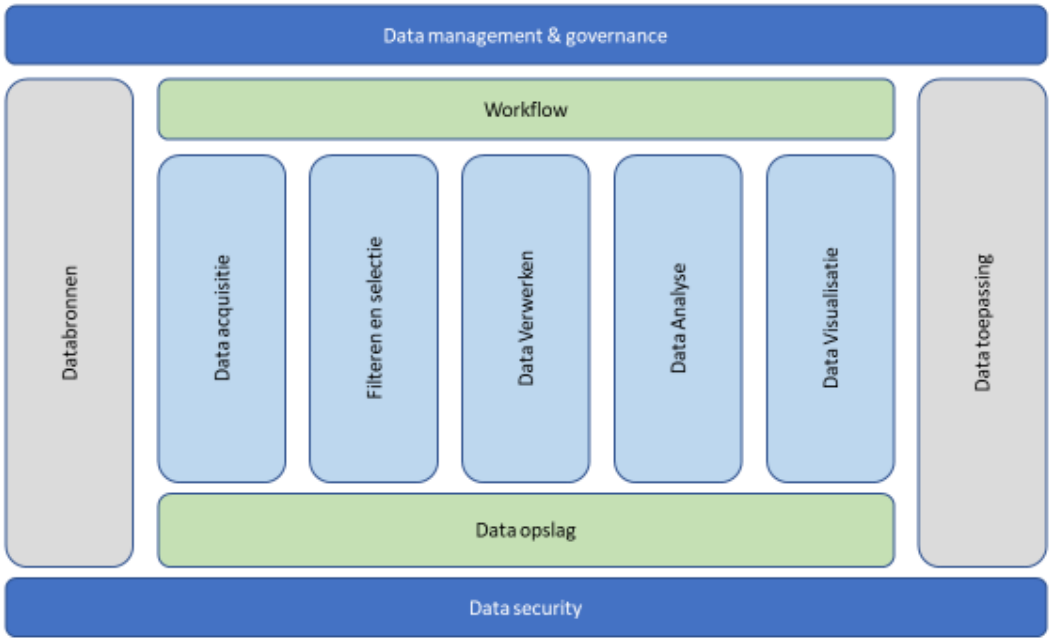
- **Gebruik**, data die geproduceerd wordt zal op zeker moment gebruikt worden, bijvoorbeeld bij het nemen van beslissingen op basis van een data analyse, alerting bij afwijkingen in de data productie of het gebruik in werkprocessen (op basis waarvan beslissingen genomen worden).
- **Integratie of transport**. Soms is er een fysieke of modelmatige afstand tussen de data die geproduceerd wordt en de data die gebruikt wordt. Denk bijvoorbeeld aan de verschillen tussen transactionele- en DWH datamodellen. Maar ook aan de plaats waar de data geproduceerd wordt (smart meter in het veld) en de plaats waar het gebruikt wordt (besturingscentrum).
- **Opslag,** Opslag maakt data persistent bijvoorbeeld wanneer er een tijdsverschil is tussen data productie en gebruik of als de data op een later tijdstip of andere context opnieuw gebruikt kan worden in een andere gebruiksvorm. Daarnaast kan opgeslagen data gebruikt worden om efficiëntie van productie en gebruik te verhogen (hergebruik van data). Reden om vanuit iedere stap in de levensloop een pijl te tonen waarbij de data beide richtingen op kan stromen (van en naar de opslag.

Bij een data levensloop wordt wel eens veronderstelt dat het ontstaan, mutatie en einde van data entiteiten weergeeft. Dat is niet het geval. Ook al is dat een interessante modelleervorm (zie bijvoorbeeld de Object Event Table) deze levensloop heeft een beperktere scope. Desgewenst is het wel mogelijk meerdere levenslopen achter elkaar te plaatsen als dit relevant is.

In de paragraaf over stakeholders wordt een voorbeeld getoond hoe de levensloop toegepast kan worden bij het toelichten van de datamodelleringsvormen

Data pipe

De data pipe is een detaillering van de datalevensloop en is wordt veel toegepast in (big) data integratie projecten. Het is feitelijk een raamwerk waarin je verschillende projectactiviteiten, deliverables en modelleervormen kunt afbeelden. Dit helpt om de complexiteit op eenvoudige wijze in kaart te brengen. Onderstaande afbeelding toont de data pipe



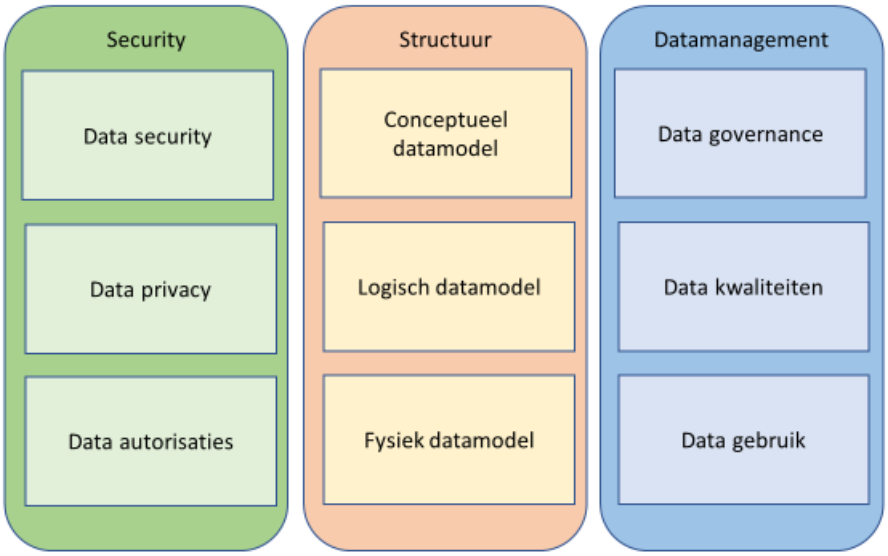
De datapipe wordt gelezen van links naar rechts en toont de stappen die genomen moeten worden om de productie van data in te kunnen zetten in een data toepassing (datagebruik). Daarnaast zijn er een aantal extra dimensie toegevoegd namelijk data management en security aspecten. In onderstaande opsomming een korte toelichting op de onderdelen:

- **Databronnen**, gegevensverzameling die gebruikt worden als grondstof voor de data toepassing
- **Data acquisitie**, activiteiten die de verkrijging van de relevante databronnen bewerkstelligen
- **Filteren en selectie**, bewerken van de gegevens vanuit de databronnen tot die datasets die relevant zijn voor de toepassing
- **Verwerken**, transformatie, manipulatie en verrijking om het datamodel geschikt te maken voor een adequate data analyse, visualisatie en toepassing
- **Analyse**, activiteiten waarbij de getransformeerde data gebruikt wordt voor het zoeken naar verbanden, patronen of statistische verhoudingen
- **Visualisatie**, zichtbaar maken van de analyse resultaten ter ondersteuning van de analisten of voor presentatie aan andere stakeholders
- **Toepassen**, inzet van data in verschillende vormen van besluitvorming
- **Dataopslag**, opslag van de data en tussenproducten voor later gebruik in vervolgstappen of andere datapipes
- **Security**, beveiligings- en privacy aspecten van data
- **Data management**, data is een asset en daarom dient er management op plaats te vinden om de waarde van data te verhogen of behouden
- **Workflow**, automatiseren en standaardiseren van bewerkingsstappen op de data

Zoals reeds genoemd is dit een detaillering van de datalevensloop en gaat dit met name in op de integratie aspecten. Dit kan betekenen dat dit raamwerk niet bij elke datamodelleervorm relevant is.

Raamwerk

Naast het meer dynamische levensloop aspect is het mogelijk om datamodelleervormen op een architectuurraamwerk af te beelden. Dit raamwerk is specifiek voor datamodellering en bestaat (zoals wel vaker) uit een drie maal drie raamwerk. Onderstaande afbeelding toont dit raamwerk.



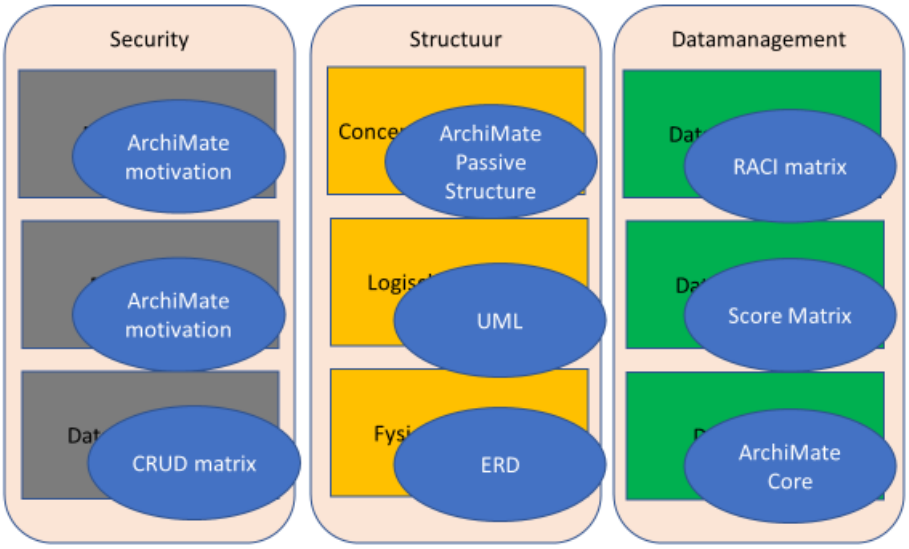
Dit raamwerk gaat uit van een drietal kolommen:

- **Structuur**, modelleren van de structuur van data zoals het opgeslagen, of toegepast wordt.
    - Fysiek datamodel, modellering van data gebaseerd op het technische platform of implementatie, bijvoorbeeld bij data opslag en transport. Dit is veelal specifiek en gebaseerd op een gekozen technologie
    - Logisch model is een platform onafhankelijk model en wordt meestal gebruikt voor een abstracte weergave van een fysiek datamodel
    - Conceptueel model is het meest abstracte model voor het beschrijven van datastructuren. Het laat veelal details van structuur weg maar gaat in op hoe data zich verhoudt tot andere concepten als bedrijfsprocessen, informatiesystemen of organisatie eenheden
- **Security**, data is de dragen van de informatie inhoud en security is daarom een belangrijk aspect ter bescherming van deze informatie
    - Data autorisaties, modellen om te beschrijven op welke wijze en voor welke persoon of toepassing data entiteiten toegankelijk zijn
    - Data privacy, privacy is een bijzondere vorm van security gericht op het beschermen van persoonsgegevens

    - Security, maatregelen en risico's rond het gebruik en met name de misbruik van gegevensverzamelingen
- **Data Management**, data is een asset voor veel organisaties en heeft daarmee waarde. Data Management zijn de activiteiten die zorgdragen dat deze waarde van data behouden of verhoogd wordt.
    - Datagebruik, op welke wijze en door welke personen en toepassingen wordt de data gebruikt
    - Datakwaliteiten, data heeft bepaalde kwaliteiten waarmee de waarde van data bepaald kan worden. Op basis hiervan kunnen data verhogende maatregelen genomen worden
    - Data governance, inrichting van de organisatie ten behoeve van het bewaken van de kenmerken die data tot een asset maken voor de organisatie

Het model is weergegeven als een drie bij drie matrix. Feitelijk zijn in dit model ook een drietal lagen te definieren. De indeling van de concepten in de kolommen houdt hier reeds rekening mee. Echter de dimensies van de lagen in de verschillende kolommen zijn verschillend vandaar dat er geen lagen als bedrijfs- en applicatie laag in deze situatie niet toepasbaar zijn. We gebruiken daarom de lagen wel maar beelden die niet af in het raamwerk.

Dit raamwerk kan goed gebruikt worden om de verschillende modelleervormen tot elkaar in verhouding te brengen. Onderstaande afbeelding toont op welke wijze dit gedaan kan worden. Dit voorbeeld toont slechts een deel van de modelleervormen maar geeft een goed beeld van het toepassen van het raamwerk.

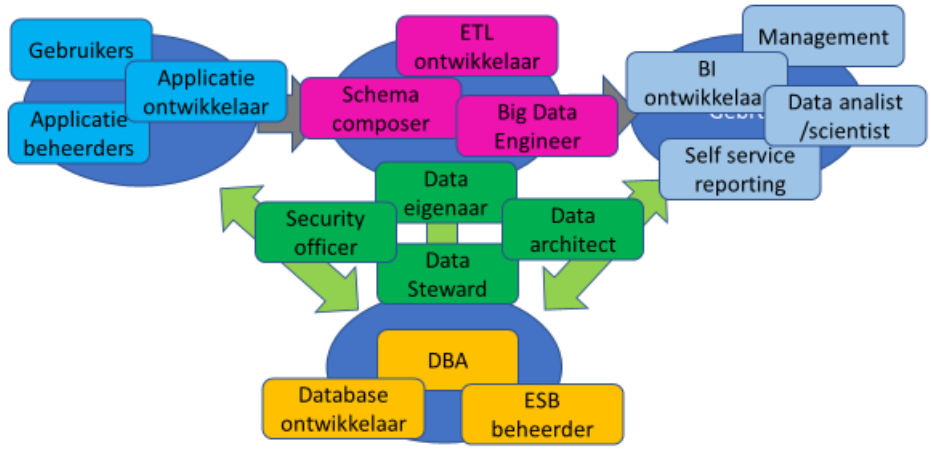


Stakeholders

Opstellen van (data)modellen doe je meestal niet zonder reden. Veelal is het gericht om betrokkenen of stakeholdes te informeren, te overtuigen of besluiten te laten nemen over een bepaald aspect van het toepassen van data. Dit toepassen kan vanzelfsprekend zijn op vele gezichtspunten betrekking hebben. Denk hierbij bijvoorbeeld aan data transport en integratie, opslag maar ook op het vlak van data toepassen in werkprocessen of informatiesystemen.

Stakeholders hebben kenmerken die van invloed zijn op de keuze van de juiste modelleervorm. Bijvoorbeeld ontwikkelaars of beheerders hebben behoefte aan meer details op het fysieke of implementatievlak. Echter gebruikers of vertegenwoordigers van de business zijn meer geïnteresseerd in hoe de concepten toegepast worden in een organisatie. Reden om stakeholders in kaart te brengen en te benoemen voor de verschillende modelleerwijzen.

Stakeholders kennen veel verschijningsvormen en het maken van een indeling of raamwerk is daarom niet goed mogelijk. De eerder benoemde indelingsvormen zijn daarbij hulpmiddelen die daarbij kunnen helpen. Onderstaande afbeelding toont een voorbeeld van hoe stakeholders afgebeeld worden op de data levensloop.



Modelleerwijzen

In dieze serie whitepapers worden een groot aantal modelleervormen beschreven. Modelleervormen zijn gebaseerd op modelleerwijzen. De modelleerwijzen zijn feitelijk modelleerpatronen op basis waarvan een modelleervorm wordt uitgewerkt.

Binnen datamodelleren worden een beperkt aantal modelleerwijzen toegepast. Sommige zijn veelvoorkomend anderen worden slechts sporadisch toegepast. De belangrijkste modelleerwijzen zijn:

- **Lijsten**, weergave van data entiteiten in opsommingen en lijstweergaven. Een veel toegepaste modelleerwijze binnen modelleertools om snel overzicht te krijgen van de beschikbare data entiteiten. Binnen modelleervormen zijn de lijsten weinig toegepast omdat er aan de vorm weinig toegevoegd kan worden vanuit modelleerperspectief
- **Bomen**, weergave van data entiteiten in een hiërarchische boomstructuur met veelal één "root" of begin entiteit waar vervolgens nul, één of meerdere "child" of kind entiteiten zijn gekoppeld. Kenmerkend is dat een kind entiteit slechts één ouder entiteit heeft. Voorbeeld van een boom is de begrippenboom of thesaurus.
- **Matrices,** weergave van de relaties tussen twee soorten entiteiten in een tabel of matrix weergave. Matrix weergave is goed bruikbaar in situaties waar tussen twee unieke entiteitsoorten slechts een koppeling kan bestaan. Vaak wordt de koppeling (lees matrixcel) verrijkt met extra selecties om een vierde dimensie toe te voegen die de koppeling typeert. Denk aan de CRUD matrix.
- **Grafen,** gebaseerd op de wiskundige grafentheorie waarbij entiteiten (nodes) via associaties of koppelingen (edges) aan elkaar gekoppeld zijn. Grafennotatie maakt het mogelijk om enerzijds meerdere associaties tussen twee entiteiten te modelleren. Daarnaast worden de entiteiten en de associaties veelal verrijkt met extra notaties zoals associatietypen en de cardinaliteiten van een koppeling. Voorbeeld van een graaf modelleerwijze is het UML klasse diagram.
- **Predicaten,** is een op taal gebaseerde modelleerwijze waarbij veelal op basis van drie aan elkaar gerelateerde entiteiten de modellen opbouwen. Bijvoorbeeld aan het predicaat "Student volgt Cursus" Voorbeeld van predicaatnotaties is NIAM.

Modelleervormen

In dit whitepaper hebben we een introductie gegeven van datamodellering en een aantal raamwerken geschetst op basis waarvan de verschillende datamodelleerwijzen in verband gebracht kunnen worden tot elkaar. Wat is het doel van deze exercitie?

Dit whitepaper is een eerste in een serie van whitepapers waarin we verschillende datamodelleervormen beschrijven. Dit beschrijven willen we graag op basis van een licht gestandaardiseerde doen, zodat het mogelijk wordt om op basis van deze beschrijvingen een vergelijking te maken tussen de verschillende modelleervormen. Hiermee wordt het mogelijk om op basis van de beschrijvingen en de context van de eigen situatie een beslissing te nemen. Welke modelleervorm het beste ingezet kan worden binnen de eigen context.

Op dit moment zijn de volgende modelleervormen (in alfabetische volgorde) beschreven in een whitepaper:

- ArchiMate (primair en secundair)
- Begrippenboom
- CRUD Matrix
- Data Flow Diagram
- ER Diagram
- RACI Matrix

- Score Matrix
- SIPOC
- UML klassediagram

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