

DE TRANSFORMATIE NAAR

EEN DATA-GEDREVEN ORGANISATIE IS GEEN ILLUSIE !

4E EDITIE

LAILA FETTAH
RONALD MEIJER
JAN SCHRAVESANDE

BLAUWDRUK VOOR EEN VEERKRACHTIGE DATA-GEDREVEN ORGANISATIE

VOORWOORD

IBM zet zich in om klanten en business partners succesvol te maken in de transformatie naar een data-gedreven organisatie. Het is onze ambitie om als kennispartner relevant te zijn voor onze opdrachtgevers en business partners en de strategie van bedrijven te ondersteunen. Data-gedreven organisaties maken veelal gebruik van Hybrid Cloud om data te ontsluiten en AI om deze data te verwerken en te gebruiken.

De strategie van IBM is Hybrid Cloud en AI. Deze helpt om de digitale transformatie van bedrijven te versnellen en succesvol te zijn in het bouwen van strategische platformen en ecosystemen.

Dit boek is het resultaat van focus, ambitie en inzet van architecten uit het IBM-overheidsteam. Samen werken aan een boek over data-gedreven organisaties geeft enorm veel positieve energie.

Het schrijven van een boek behoort niet tot de dagelijkse werkzaamheden van onze architecten. Het schrijven van boeken is geen voorwaarde om IBM-architect te worden. Ik ben echter van mening dat ze er goed in geslaagd zijn om een interessant en zeer leesbaar boek te schrijven!

Pas op! Dit is een 'lees- en werkboek', het is geen 'academisch studie- en leerboek'.

Hopelijk leest u dit boek met belangstelling en plezier. Het is geschreven met passie voor het onderwerp en de wens u te helpen op uw reis naar een data-gedreven organisatie.

Johan Heij
IBM Nederland

INTRODUCTIE

In 2021 produceerde het CIO Office van de Nederlandse overheid een strategisch document genaamd “I-strategie Rijk”¹. Deze strategie bevatte een zeer interessant thema, namelijk iets wat we in het Nederlands ‘informatiehuishouding’ (IHH) noemen. De normale Engelse term hiervoor is information management. Hoewel ‘informatiehuishouding’ officieel niet bestaat, gebruiken we deze term omdat deze precies de boodschap weergeeft die we willen overbrengen: niemand, nou ja bijna niemand, houdt sowieso van huishouden! Waarom zouden mensen het huishouden in hun informatie domein dan wel leuk vinden? Alleen maar om het netjes en opgeruimd te houden?

Wij hebben een bekende Engelse uitdrukking omarmt: *“A place for everything and everything in its place”!* Dit thema in de I-strategie wekte onze nieuwsgierigheid en tot op de dag van vandaag beheerst dit onderwerp onze activiteiten, onze gedachten, discussies en onze ideeën. Het houdt ons van de straat!

We ontwikkelden een ‘point of view’ (wat absoluut geen visie is), maar een mening over dit onderwerp en hier hebben we met veel mensen binnen de Nederlandse overheid over gesproken. Van beleidsmakers en besliszers tot informatiearchitecten en technische mannen en vrouwen die verantwoordelijk zijn voor het beheer van Content Management Systemen en de bijbehorende informatiehuishouding.

Wij zijn er sterk van overtuigd dat een goede informatiehuishouding de basis en een absolute voorwaarde is voor een data-gedreven organisatie. Dit geloof en inzicht is de afgelopen periode door onze betrokkenheid bij dit dossier sterk

¹ <https://www.rijksoverheid.nl/documenten/beleidsnotas/2021/09/06/i-strategie-rijk-2021-2025>

gegroeid. Focus op Generatieve AI en AI-toepassingen vraagt ook veel van een goede informatiehuishouding. Eigenlijk kun je wel stellen *geen AI zonder een goede IA* (Informatie architectuur).

Onze initiële focus was gericht op ongestructureerde data die de basis vormen voor veel overheidsdossiers (met name documenten/dossiers). De manier waarop deze worden gebruikt in overheidsorganisaties, de manier waarop de levenscyclus van deze data wordt beheerd en hoe overheidsorganisaties een betere aanpak van informatiehuishouding kunnen ontwikkelen.

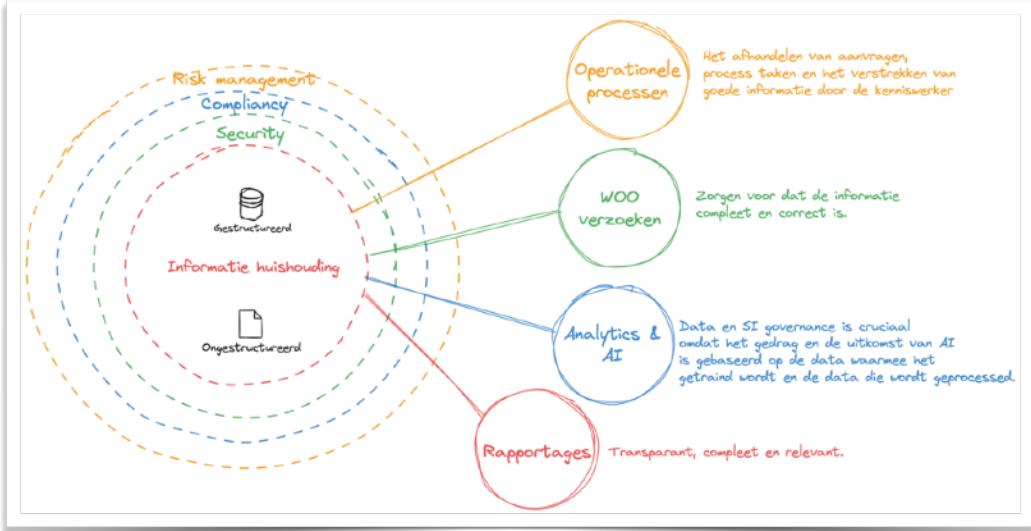
Intussen beseffen we hoe belangrijk het is om alle data in scope te hebben (gestructureerd, semi-gestructureerd en ongestructureerd) voor een goede informatiehuishouding en om een data-gedreven organisatie te worden. Om een total beeld te krijgen (360 graden overzicht) moet transactionele dat en dossier data bij elkaar gebracht worden. Van alle organisaties in de wereld is de overheid bij uitstek een voorbeeld van een organisatie dat data-gedreven wil en eigenlijk moet zijn om een goede informatiepositie te hebben om wet- en regelgeving goed te kunnen uitvoeren (zonder fouten dus)!

Om data te kunnen gebruiken en om de processen van organisaties te sturen, moeten organisaties informatie bewust worden en een nieuwe aanpak en nieuwe voorzieningen ontwikkelen.

Data speelt een cruciale rol in veel besluitvormingsprocessen. Voorwaarde is dat data voldoet aan twee belangrijke criteria:

1. *Governance* - data is onder governance en het gebruik van data in bedrijfsprocessen gebeurt op een verantwoorde manier (denk aan proportionaliteit en doelbinding).

2. Architectuur - de informatie architectuur zorgt voor kwalitatieve goede data.



Waarom informatiehuishouding?

Maak niet de fout door te denken dat het upgraden van je data-warehouse met het adopteren van wat AI-technologie voldoende is om te transformeren naar een data-gedreven organisatie. Er zijn nieuwe technologieën, processen en vakgebieden ontstaan die allemaal nodig zijn bij deze reis. Dit alles brengt ook nieuwe uitdagingen met zich mee.

Het is tegenwoordig bijna een modewoord of zelfs een hype:

“Wij willen een data-gedreven organisatie worden”. Daarop zeggen wij: “ja, natuurlijk wil je dat!”. Dat is bijna hetzelfde als roepen “Ik wil een moderne zelfrijdende auto!, terwijl je zelf nog in



een DAF-33 rijdt”.

We begrijpen ook wel dat je niet de Fred Flintstone van Bedrock wil zijn....., maar om data-gedreven te worden moet je wel enkele (r)evolutionaire stappen maken.

Dit boek geeft je een overzicht van diensten, architectuur patronen en methoden die je kunnen helpen bij het implementeren van een data-gedreven organisatie. Naast dit overzicht biedt het een reeks best-practices om IT-organisatie te helpen valkuilen te vermijden.

De meeste gepubliceerde artikelen over data en AI beperken zich tot vraagstukken binnen de kaders van het data domein. Dat betekent dat aspecten als operationalisering, veiligheid en voortbrenging vaak buiten scope van dit domein liggen. Vraagstukken als cultuur, complexiteit en best-practices worden veelal goed beschreven, maar zijn gericht op het data domein zelf en op (natuurlijk) interessante aspecten als machine learning, je weet wel, die gadgets en hypes van vandaag de dag. Maar als je succesvol wil zijn met data-oplossingen moet je ook over de grenzen van het data domein kijken en juist al die aspecten bij elkaar brengen.

Met de inhoud van dit boek is het mogelijk een holistisch overzicht te creëren van de benodigde (IT)diensten, door gebruik te maken van architectuur patronen en ontwikkel methoden.

Je kunt het vergelijken met koken: we beschrijven zowel het recept met ingrediënten (de (IT)diensten) en de aanpak om de maaltijd te bereiden. We hebben de diensten vrij gedetailleerd beschreven en we denken dat het dit boek erg nuttig maakt als naslagwerk.

De aanpak (architectuur patronen en ontwikkel methoden) is te vergelijken met de bereidingswijze: je kookt, braadt, bakt of je gebruikt andere manieren om je maaltijd te bereiden.

De architectuur patronen en methoden beschrijven we in het

tweede gedeelte van het boek.

Omdat deze soms abstract en theoretisch zijn proberen we deze in Jip en Janneke taal uit te leggen en met eenvoudige voorbeelden te omlijsten. Dat doen we niet in detail, maar we geven een overzicht van de meest bekende raamwerken en methoden en leggen daarbij uit waarom dat specifieke raamwerk of aanpak belangrijk kan zijn om jouw data-gedreven organisatie op te bouwen. Als je meer wilt weten over die aanpak of dat raamwerk, verwijzen we je naar specifieke artikelen hierover.

Door architectuur patronen en raamwerken uit te leggen in de context van een data-gedreven architectuur, vertrouwen we erop dat je de keuzes kunt maken die passen bij jouw specifieke context, cultuur, doel en wensen.

Het idee is dat je begint met het hoofdstuk één, (daarom hebben we het ook hoofdstuk één genoemd!). Dit hoofdstuk geeft je een overzicht en een algemeen idee van wat een data-gedreven organisatie is. De overige hoofdstukken kun je willekeurig raadplegen, zolang je deze maar niet vergeet te lezen!

Als je een veerkrachtige data-gedreven organisatie wilt worden, heb je het complete verhaal nodig. Wij hanteren het principe: schrijf het zo simpel mogelijk op. Wij realiseren ons heel goed dat dit ten koste kan gaan van nuances of details. We waren echter van mening dat het begrijpen van het grote geheel, de context en het begrijpen van de relaties tussen architecturdomeinen, organisaties gaat helpen met hun reis.

Dit boek is daarom ook bedoeld voor solution architecten, enterprise architecten, business- en IT-consultants en strategen. Ook voor CIO's, CTO's, CDO's en programmamanagers kan het waardevol zijn.

Bij dit boek zit een poster (A0 afmeting) met een generieke indeling van de blauwdruk. We hebben een hoofdstuk toegevoegd waarin wordt beschreven hoe je deze poster kunt

Dit is onze missie voor jou:

Ontwikkel deze blauwdruk, maak hem eigen en gebruik hem
in je discussies met jouw stakeholders.
maak deze blauwdruk tot een levend bezit voor jou en je
organisatie.

gebruiken om je eigen blauwdruk te ontwikkelen door use-cases
te projecteren op de verschillende services.

Het laatste woord in deze inleiding gaat over transparantie. Zakelijke transparantie is (volgens Forbes²) het proces van open, eerlijk en duidelijk zijn over verschillende zakelijke activiteiten. Transparante bedrijven delen informatie met betrekking tot prestaties, bedrijfsopbrengsten, interne processen, inkoop, prijzen en bedrijfswaarden. Je bent verplicht om alle belanghebbenden te informeren met nauwkeurige informatie om gerespecteerd te worden als een transparante en duurzame onderneming. Om deze rapportage te kunnen doen, moet je over nauwkeurige gegevens beschikken. Dus ook dat is een reden om de informatiehuishouding op orde te hebben.

Geniet van dit boek!

Laila Fettah, Ronald Meijer, Jan Schravesande

² <https://www.forbes.com/sites/mikekappel/2019/04/03/transparency-in-business-5-ways-to-build-trust/>

DE TWEEDE, DERDE EN VIERDE EDITIE

In de tweede editie van ons boek hebben we twee onderwerpen toegevoegd op verzoek van onze lezers.

De eerste vraag was betreffende het gebruik van het boek. In de workshops die we hebben gehouden hebben we diverse invalshoeken gezien over het gebruik. Deze lessen hebben we toegevoegd aan dit boek.

Als tweede hebben we het onderwerp over duurzaamheid verder uitgebred. Wanneer we kijken naar data en kunstmatige intelligentie (AI) zijn dat grootgebruikers voor wat betreft energie. We voelde ons verantwoordelijk om dit onderwerp bij onze lezers onder de aandacht te brengen.

Uiteindelijk is die verantwoordelijkheid uit de hand gelopen en hebben we zelfs een heel boek aan dit onderwerp geweid. Daarom heet dit boek uit de oranje serie nu deel 1, er bestaat dus een deel 2 die gaat over duurzaamheid.

Hoewel we ons boek tientallen malen doorzocht hebben op fouten, waren er toch een paar zeer oplettende lezers die een fout hadden gevonden. Deze hebben we aangepast. Mocht je er desondanks toch nog één vinden dan hebben we die er waarschijnlijk opzettelijk in gelaten.

Door de vele discussies, vergaderingen en workshops hebben we ook veel ideeën aangereikt gekregen. Dit heeft ertoe geleid dat we in de derde druk drie onderwerpen hebben toegevoegd. De

eerste was onvermijdelijk, namelijk generative AI (genAI). De opkomst van ChatGPT is alleen mogelijk door het gebruik van astronomische hoeveelheden data om dit AI-model te trainen. Het tweede onderwerp dat we hebben toegevoegd is data-gedreven programmeren. Wat heeft dat met een data gestuurde organisatie te maken vraag je je misschien af? Nou lees het hoofdstuk! En als derde hebben we de data catalogus toegevoegd. Een belangrijk instrument om governance op data te kunnen uitoefenen.

Ten slotte deze vierde uitgave. Een vertaling in het Nederlands. Meer dan dat! Door de opkomst van genAI hebben we informatie toegevoegd over data en AI governance. Daarnaast vind je nog een typisch Nederlands voorbeeld over de haven van Rotterdam. Geen letterlijke vertaling maar echt een vierde druk omdat wij ervaren dat een (IT) boek in het Nederlands toch wel een stuk makkelijker wegleest. Verder vertellen we iets over de achtergrond van het ontstaan van dit boek en dat is gerelateerd aan de Nederlandse politiek en de IT problematiek. Ten slotte, niet onbelangrijk, merken we dat er te weinig aandacht is voor data security.

Security moet je niet achteraf doen, maar meenemen vanaf het allereerste moment! Dus we hebben hier aan aantal bladzijden aan geweid. Quantum speelt ook een belangrijke rol in het beveiligen van informatie. We hebben hier een kort stukje over toegevoegd.

Hoe organiseer je een goede data governance? Wij denken dat het DMBoK framework hierbij goed kan helpen.

Als laatste: In de eerste editie van het boek stond een disclaimer dat wij Nederlandse gedachte in het Engels hebben geprobeerd te vertalen. Wij stonden nu voor de uitdaging om deze (af en toe) kromme Engelse uitdrukkingen weer in begrijpelijk Nederlands terug te vertalen. Pfffff.

Table of Contents

Voorwoord	3
Introductie	4
De tweede, derde en vierde editie	10
De oorsprong van onze zienswijze	15
I. De Context	19
1. Een data-gedreven organisatie	20
1.1. Een veerkrachtige organisatie	22
1.2. ‘Verschillende soorten systemen’	23
2. Hoe kan ik dit boek gebruiken?	28
3. Architectuur context	32
3.1. Geen artificial intelligence zonder informatie architectuur	32
3.2. Architectuur overzicht voor een data-gedreven organisatie	33
3.3. Frameworks (raamwerken)	38
3.4. People	39
3.5. Kan AI de kwaliteit van bedrijfsprocessen verhogen?	40
3.6. Hybrid cloud	41
3.7. Data repatriëring als een serieuze waarschuwing	42
3.8. RealTime Analytics	43
3.9. Smart port: Havenbedrijf rotterdammer en internet of things	44
3.10. Roadmap en planning	46
II. De Blauwdruk	47
4. Structuur van de blauwdruk	48
5. Business Processes	54
6. Business apps (SoR)	58
7. Business process support Services	59
7.1. business process management (BPM) en business process monitoring (BPM)	64
8. Data zones	66
9. Integration Services	68
9.1. transport- of toegangstechnologie selecteren	73
9.2. Het ophalen van gegevens	74
9.3. het verstrekken van gegevens	76
10. Sorts of Data	78

11. Data Analysis Services	83
11.1. AI governance Staat haaks op het agile manifest	87
11.2. Het algoritmeregister	90
11.3. Verantwoord algoritmegebruik	91
11.4. Doelen van het Algoritmeregister	91
12. Content analysis services	94
12.2. Collaboratie	103
12.3. Records Management	107
13. Data governance and compliance Services	112
13.1. Data governance versus ai governance	113
13.2. Een Data Catalogus, hoe moeilijk kan het zijn?	117
14. CI/CD Pipeline services	121
14.1. Devops versus MLOPS	124
15. Hybrid cloud management Services	125
15.1. Application Runtime	128
16. Platform Hosting Services	130
16.1. Analytical werklasten	132
16.2. Hosting diensten in context van analytics	132
16.3. Storage	135
16.4. Hyper-converged infrastructuur	137
16.5. Quantum als hosting platform	137
17. Data Security Services	139
17.1. Stel je eens voor.....	142
17.2. De weakest link	146
17.3. Classificeren van data	147
17.4. De Quantum bedreiging	150
III. Processen en Raamwerken	157
18. Duurzaamheid	158
18.1. Responsible code	159
18.2. Responsible Infrastructure	160
18.3. Responsible Data Center	161
18.4. Responsible Data Usage	162
18.5. Responsible Systems	162
18.6. Responsible Impact	163
18.7. Responsible Computing Values	163
19. Raamwerken	165
19.1. Busops	165
19.2. Devops	166
19.3. ModelOps	168
19.4. Aiops	170

19.5. Garage	172
20. Data Mesh, het antwoord op veerkracht?!	174
21. Beheer van het Data Platform met behulp van DevOps en SRE teams	
180	
21.1. Verschillende Systemen Samenbrengen	181
22. Data Fabric	183
23. Vertrouw jij op GenAI om beslissingen voor jou te nemen?	188
23.1. LLM Voorbeelden	189
23.2. Large Language Models in een Data-Gedreven context	191
24. Data-Driven Programming	192
25. En verder	196
25.1. Aan de Slag	198
25.2. Notities	198
Epiloog	202
Appendix	205
Glossary	207
Over de schrijvers	211

DE OORSPRONG VAN ONZE ZIENSWIJZE

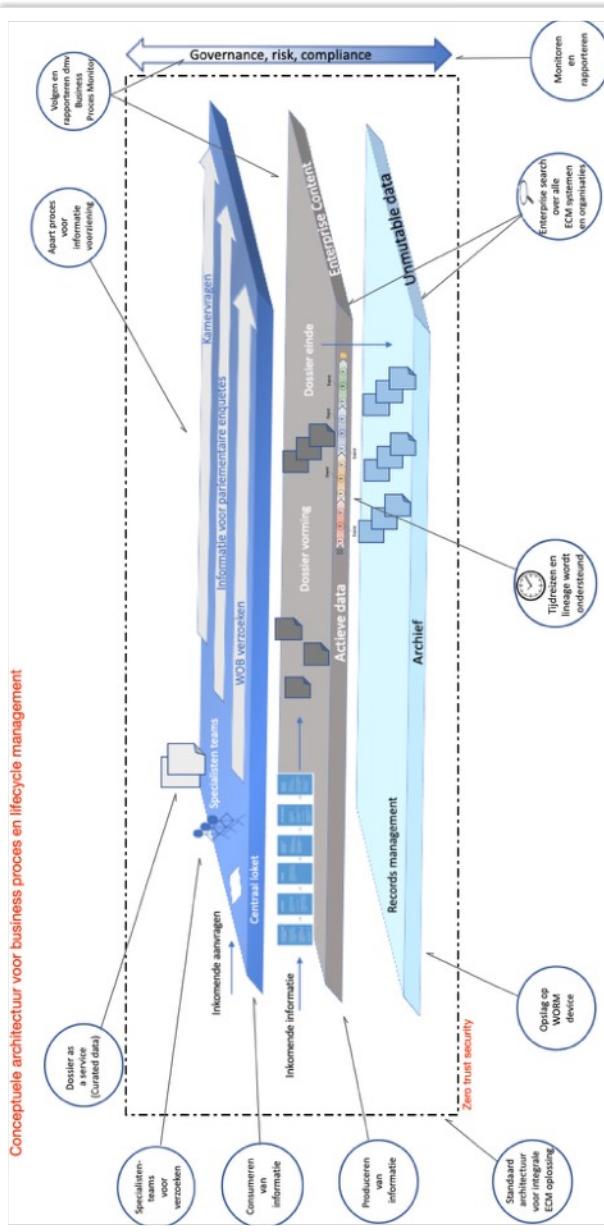
Dit boek is ontstaan vanuit onze zienswijze op het IHH thema van de overheid. In deze zienswijze hebben wij op basis van veel gesprekken met organisaties in de overheid een weergave gegeven van de verschillende problemen die er zijn op het gebied van informatiehuishouding in relatie tot ongestructureerde data (content). Laten we hier een korte opsomming geven:

- Gebrek aan (proces) standaardisatie. Iedere organisatie hanteert ‘eigen’ standaarden op het gebied van content processing.
- Gebrek aan (technische) standaarden. DocDirect levert een generieke voorziening voor content management maar daar maken maar weinig organisatie gebruik van. Er is geen consensus over een architectuur die overheidsbreed is.
- WOO (Wet Open Overheid, voorheen Wet Openbaarheid van Bestuur (WoB) verzoeken behandelen is zeer specialistisch werk en vergt veel tijd.
- Er is geen goed beheer over de levenscyclus van content. Hierdoor voldoen veel organisatie niet aan de archiefwet.
- Veel content valt niet onder ‘beheer’ van het content management systeem. Mail, fileshares en andere omgevingen bevatten veel relevante content.
- Tijdreizen is niet goed geïmplementeerd.
- ...

Het gaat hier dus echt om de drie dimensies People, Process en Technology:

People: Hoe gaan mensen om met content? Doen ze dit netjes binnen het domein van content management of wordt het voorlopig maar even ‘ergens’ geparkeerd?

Process: Wordt het acquireren, creëren van content op eenduidige wijze gedaan of hanteren organisaties hun eigen



Informatie Huishouding

methodes?

Technology: is er sprake van een federatief content management systeem of is het een organisatie specifieke inrichting van de infrastructuur? Zijn er standaarden of is er onderlinge competitie?

Bovenstaande conceptuele plaat geeft een overzicht van de processen voor content management. In een plaatje lijkt het simpel maar dat is het natuurlijk niet!

Deze plaat is een onderdeel van de zienswijze die wij hebben gedeeld met veel organisaties in de overheid. Bij een van deze sessies was de Rijkscommissaris voor Informatiehuishouding aanwezig: Arre Zuurmond. Hij legde in deze sessie zijn visie op informatiehuishouding uit. Enerzijds over standaardisatie van processen door het hanteren van zgn. Archetypen. Iedere archetype zou een standaard proces moeten omarmen!

Daarnaast, en hier komt het, de relatie met gestructureerde data.

De informatiepositie van de overheid en de informatiepositie van burgers zijn van cruciaal belang voor het vertrouwen maar ook voor innovatie in wet- en regelgeving.

Met diverse voorbeelden heeft hij aangegeven waarom data-gedreven alleen werkt als de informatiepositie van zowel overheid als burger op orde is. Tijdigheid van informatie speelt hierbij een grote rol. Om besluiten te kunnen nemen over dossiers (toeslagen, belastingen, uitkeringen, enz) dient de informatie volledig en tijdig te zijn. Niet kijken in de data van gisteren om een besluit vandaag te nemen. Alleen op basis van actuele en correcte data voorkom je (grote) fouten bij het beoordelen van dossiers.

Deze inzichten waren voor ons de aanleiding om ons onderzoeksgebied uit te breiden naar alle data. Gestruktueerd, semi-gestruktueerd en ongestruktueerde data!
Data-gedreven gaat dus om informatiepositie:

I. DE CONTEXT

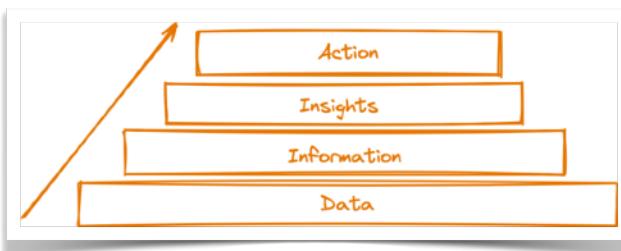
“Als je niet weet waar je staat, wordt het lastig om te bepalen waar je heen moet”

1. EEN DATA-GEDREVEN ORGANISATIE

Laten we het eerst eens hebben over een paar definities zonder te vervallen in een academisch woordenspel. Om te beginnen, waarom willen organisaties data-gedreven worden?

Om echt een data-gedreven organisatie te zijn, betekent dit dat je data ziet als kostbare diamanten. Je onderkent de kracht van informatie om deze te gebruiken, om te kunnen anticiperen op nieuwe situaties (predictive) en eventueel om je bedrijfsvoering (bij) te sturen (prescriptive): wat gaat er gebeuren en welke actie kunnen we nemen? Dat kan betekenen dat we burgers beter ondersteunen bij vragen over subsidies of het oplossen van inefficiënties in processen, of....het winnen van Formule 1-races.

Het komt allemaal neer op het transformeren van de data grondstof naar bruikbare informatie die leiden tot zinvolle inzichten voor je bedrijf om op te sturen. Dit betekent dat het van cruciaal belang is om je datastrategie af te stemmen op je bedrijfsstrategie. Om daar te komen, moeten we elke keer een stap hogerop in de piramide zetten. Om van data naar informatie te gaan moeten we context aan de data toevoegen. We komen van informatie naar inzichten door logica en intelligentie toe te voegen en vanuit die inzichten kunnen we acties genereren.



Value of Data

Als voorbeeld:

wat dacht je van een organisatie dat data-gedreven ademt en letterlijk tot het uiterste gaat (qua snelheid dan): Formule 1-racen! Tijdens elke race genereren 120 sensoren in elke auto 3 GB aan data, elke seconde worden 1.500 datapunten uitgelezen³. Ze rijden letterlijk op data.

In het weekend van 29 mei 2022 stond het belang van data en inzicht voorop tijdens de Grand Prix van Monaco. Met Ferrari als favoriet op pole position hadden zij eigenlijk positie 1 en 2 moeten halen en Red Bull plek 3 en 4⁴.

Daarmee zou Ferrari 43 punten moeten scoren en Red Bull 27. Elke punt staat voor zo'n slordige miljoen euro. Ferrari had dus 16 meer punten moeten "verzamelen" dan Red Bull. Tijdens de race veranderde het weer en daardoor waren en plotseling andere banden nodig.....

Op basis van de verzamelde data berekende Red Bull dat de afstand tussen de auto's net voldoende was om in één ronde de banden van beide auto's te wisselen.

Ferrari had dit echter niet correct berekend en dit zorgde voor verwarring en frustratie wat leidde tot vertraging bij de wissel. Toen ze terugkeerden op de baan, werden ze gepasseerd door de beide Red Bulls. Het eindresultaat was 40 punten voor Red Bull en 30 voor Ferrari. Dus een verschil van 13 miljoen ten opzichte van de verwachting en erger nog, 13 punten voor de directe concurrent.

De credits gingen naar de data scientist die de berekeningen uitvoerde; zij verdiende die dag letterlijk miljoenen dollars.

³ <https://aws.amazon.com/solutions/case-studies/formula-one/>

⁴ https://en.wikipedia.org/wiki/2022_Monaco_Grand_Prix#:~:text=In%20the%20final%20laps%2C%20Pérez,with%20Ocon%20falling%20to%202th.

11. EEN VEERKRACHTIGE ORGANISATIE

Waarom hebben we het woord veerkrachtig (in de subtitel) toegevoegd? Als lezer denk je misschien dat een data-gedreven organisatie per definitie veerkrachtig is.

Wij zijn van mening dat dit niet het geval is! Het risico van het opbouwen van een data-gedreven organisatie op basis van de situatie van vandaag en gisteren kan leiden tot verkeerde zakelijke beslissingen. Wanneer je niet gemakkelijk kunt aanpassen aan veranderingen in bedrijfsmodellen en externe gebeurtenissen, kunnen je voorspellingen van de bedrijfsresultaten mislukken. In de Formule 1 kunnen veranderingen aan auto's betekenen dat je als data scientist snel je model moet aanpassen om de juiste uitkomst te kunnen voorspellen.

Daarom moet je rekening houden met de volgende aspecten:

- Wendbaarheid; nieuwe regels en algoritmen moeten in korte tijd worden ontwikkeld en geïmplementeerd in de bestaande of nieuwe bedrijfsprocessen.
- Veiligheid en conformiteit; je moet voldoen aan wet- en regelgeving.
- Hoge beschikbaarheid; de systemen moeten beschikbaar zijn om toegang te verlenen tot de gegevens.
- Betrouwbaarheid; inzicht moet betrouwbaar en verklaarbaar zijn.
- Tijdigheid; inzicht moet op het juiste moment beschikbaar zijn. Niet alles hoeft actueel te zijn, maar je moet wel inzicht hebben in de feitelijke situatie wanneer dat nodig is.
- Schaalbaarheid; je moet kunnen opschalen bij pieken en afschalen in tijden van rust (niet de siësta maar ten tijde van minder gebruik en dat zou zomaar eens tijdens de siësta kunnen zijn).

1.2. 'VERSCHILLENDEN SOORTEN SYSTEMEN'

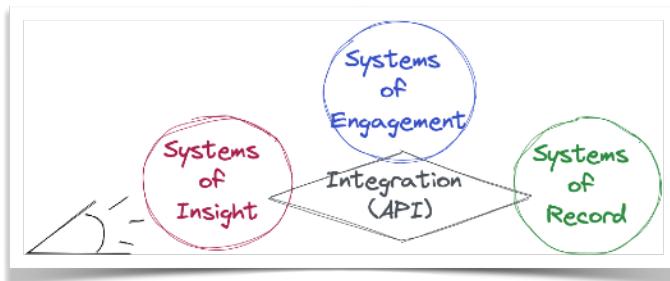
Als we naar het landschap van systemen kijken, moeten we onderscheid maken tussen onze kernsystemen, onze systemen die zijn gemaakt voor interactie zoals apps, web en portal en systemen die zijn ontworpen voor inzichten en rapportage.

De kernsystemen noemen we Systems of Record. Dit zijn doorgaans de belangrijkste administratieve- en transactie verwerkende systemen. Voor banken is dat het betalingssysteem, voor de logistiek is dat een planning en order management systeem.

Door innovatie liefhebbers worden deze systemen vaak 'legacy' genoemd. Het is echter nog steeds de kern van de bedrijfsvoering!

De Systems of Engagement zijn systemen die communiceren met eindgebruikers en interne medewerkers. Systems of Insight zijn typische data-warehouse- en rapportagesystemen.

In deze afbeelding zie je een model dat meer dan tien jaar geleden door IBM-research is gecreëerd.



Verschillende soorten systemen

Het is erg nuttig om verschillende soorten systemen te onderscheiden omdat ze allemaal hun eigen 'hartslag' hebben,

hun eigen snelheid. Gartner noemt dit 'Bi-modal IT⁵'. Wij koppelen deze verschillende systemen door middel van integratietechnologie, bij voorkeur middels API's.

Als we het over een data-gedreven organisatie hebben, ligt onze focus op de *Systems of Insight*.

Maar hoe ziet de architectuur voor zo'n systeem eruit?



⁵ <https://www.gartner.com/en/information-technology/glossary/bimodal>

In dit boek vind je ook enkele kritische opmerkingen over de mogelijke impact die Systems of Insight kunnen hebben op je IT-strategie.

Het lijkt erop dat de (technische) IT-strategie grotendeels wordt bepaald door Systems of Engagement, maar deze strategie sluit niet altijd aan bij de strategie voor de Systems of Insight. Systems of Engagement richt zich op interactie met gebruikers en stelt heel andere eisen aan het systeem. Denk na over keuzes voor de adoptie van een Hybride Cloud strategie en een ‘run anywhere’-aanpak. Dit heeft impact op je data!

Het kan betekenen dat je de strategie moet heroverwegen en niet blindelings alle trends moet volgen. Ons doel is om je over deze onderwerpen te laten nadenken voordat je achter de hypes van vandaag de dag aanrent.

De volgende tabel is een poging om de verschillen in focus op de karakteristieken voor de “soorten systemen” te benadrukken. Dit is niet zwart-wit, integendeel, je zult merken dat de grenzen tussen de systemen vervagen. Veel van de hier gebruikte termen zijn afgeleid van de ISO/IEC 25010-norm voor de kwaliteit van softwareproducten, maar het gaat niet om de exacte termen. Het zou je wat meer gevoel moeten geven over de focus op requirements⁶.

⁶ <https://iso25000.com/index.php/en/iso-25000-standards/iso-25010>.

	Systems of Record	Systems of Engagement	Systems of Insight
Positionering	Systemen met een lange levensduur. 100% accuraat.	Interactie en interface met de SoR	Downstream systemen, gegevens afgeleid van SoR en ook SoE
Users	Expert users	Klanten en medewerkers	Medewerkers en beperkte ontsluiting voor klanten en toezichthouders
Concepten	Business logic, rules, relationele databases met integriteit.	Caching and Queuing, AVG	Storage, ((No) SQL) Databases, Data Models, Data Product
Compliance	Wetgeving	Privacy, AVG	Ethische normen
Functionele toepassing	Correct en compleet	Doet wat het moet (MVP)	Doet wat het moet
Technologie	Commercial off the Shelf (COTS): maatwerk met Compiled code. Interface op maat.	Portals: Multi-channel, multi-device. Scripting. API's, Open Source. Internet.	Dashboards: Data analysis, BI en Reporting. ETL en file transfer. Big Data tot en met Cognitive analytics
Performance Efficiency	Resource Utilisation, Latency en Throughput, Sustainability	Response time	Capacity
Security	Accountability	Toegangs beveiliging	Confidentieel en Integer
Onderhoudbaarheid	Reusability, Modularity, Analyability	Modifiability, Testability	Lineage
Usability	Operationeel	User Error Protection, User Interface Aesthetics	Accessibility of data

Focus	Stabiliteit	Portability voor verschillende devices en agile veranderende user demand te ondersteunen	Data Volumes, data structures en data formats, data governance
Karakteristieken	Transactioneel gevoelig, commits en rollbacks.	Van informatie voorziening tot single user interaction.	Governance
High Availability	Data synchronisatie voor Stateful applications en transactions.	Application beschikbaarheid voor Stateless applications	Asynchronous copies om de omgeving opnieuw op te bouwen
Disaster Recovery	RPO = 0 or near 0 (two phase commit), RTO = minutes to hours	RTO = 0, Master-Slave, concept of immutability.	High Availability is de solution voor Disaster Recovery.
Reliability	Maturity, Recoverability	Fault Tolerance	Data Quality
Critical Resource	I/O en memory Bound	CPU en memory Bound	Storage en GPU bound
Schaalbaarheid	Verticaal	Horizontaal	Verticaal en Horizontaal
Framework	Iterative, Ops	DevOps, SRE	ModelOps, SRE
Trends	Modernization. Verschuiving naar packaged en standard products	Toegang tot SoR via API's. Low Code.	AI, ML, data-driven.
Cloud affiniteit	IaaS / SaaS (Private or Public)	PaaS (Public)	Data Fabric (Private)

Zoals je ziet zijn er verschillen in de eisen voor SoR, Sol en SoE (soort systemen). Behandel ze daarom niet gelijk!

Je moet je realiseren dat 80% van de software ontwikkelkosten verband houdt met kwaliteitsaspecten. Om je oplossing betaalbaar te maken, moet je een afweging maken tussen de verschillende kwaliteitsaspecten en beslissen waar je geld in investeert. Je kunt het namelijk maar één keer uitgeven!

2. HOE KAN IK DIT BOEK GEBRUIKEN?

It was de eerste vraag die ons werd gesteld op het moment dat wij de eerste versie van dit boek introduceerde. Gebaseerd op discussies, feedback en workshops zien wij een aantal manieren waarop je het boek met de bijbehorende blauwdruk kan toepassen:



Blauwdruk relevantie

1. Dit boek is een holistische, high-level introductie op het onderwerp “data-gedreven”. Het geeft inzicht in de relaties tussen de verschillende logische architectuur domeinen die wij in dit boek geïdentificeerd hebben. De samenhang tussen deze domeinen is erg belangrijke met name voor mensen met een overkoepelende verantwoordelijkheid. Bijvoorbeeld voor een CTO of Programma Manager. Je houdt hiermee het overzicht over het totaalplaatje en je kunt hiermee controleren of de juiste services worden gebruikt voor een ‘Solution’. Zo voorkom je bijvoorbeeld dubbele functionaliteit of functies die op een onlogische plaats in de architectuur worden gebouwd. In dit boek worden heel veel verschillende onderwerpen besproken. Sommige onderwerpen zijn ‘bekend’ en andere

onderwerpen zijn nieuw voor de organisatie. Om de juiste keuzes te maken heb je een compleet beeld nodig. Misschien gaat het voor de een wat te ver, maar voor anderen zal het helpen om het overzicht te bewaren en de complexe relaties tussen diverse aspecten te bewaken om een goede data-gedreven organisatie te bouwen.

2. Omdat wij hier “slechts” onze mening weergeven durven we een aantal controversiële statements te maken. We zijn geneigd om lastige vraagstukken voor ons uit te schuiven omdat deze het allemaal maar moeilijker maken. Nadenken en implementeren van de juiste werkwijze, toekennen van verantwoordelijkheden, inrichten van governance om er maar een paar te noemen. Wij durven trends ter discussie te stellen omdat we zien dat deze soms blind en ongenuineerd worden overgenomen. Wij doen dat natuurlijk niet!
Onze ambities is om deze zaken *open en bloot* op tafel te leggen zoals we dat zo mooi zeggen.
3. Als je (nog) geen referentiearchitectuur hebt is dit een super goed startpunt, althans dat vinden wij.
4. Wanneer je een referentie architectuur hebt kun je deze blauwdruk gebruiken om jouw architectuur te valideren. Wellicht heb jij services niet geïdentificeerd die wij wel geïdentificeerd hebben. Stel jezelf de vraag waarom je deze niet hebt. Misschien zijn er services die jij wel hebt en wij niet. Stel je dan de vraag of je jouw service met de juiste granulariteit hebt gedefinieerd. Misschien kun je het idee van de A0 poster als communicatiemiddel gebruiken. Wij krijgen daar namelijk veel positieve reacties op!
5. Sommige mensen hebben gevraagd of deze architectuur blauwdruk de magische versneller is om een data-gedreven organisatie te bouwen. Daar kunnen we een eenvoudig

antwoord op geven. Natuurlijk is dat het!

Uiteraard ligt dat genuanceerder, want je zult de materie die we in dit boek beschrijven je eigen moeten maken en het plaatsen in de context van jouw organisatie. Je zult (architectuur) beslissingen moeten nemen. Misschien wel over onderwerpen waar je helemaal nog niet over nagedacht had. Dat is een behoorlijke hoeveelheid werk maar absoluut de moeite waard omdat je hiermee een holistisch overzicht creëert voor alle aspecten van een data-gedreven organisatie.

Nog één belangrijk dingetje! Ga dit niet inzetten als een big bang aanpak. Dat komt er op neer dat je de oceaan aan het koken bent. Daarom hebben we het concept van use-cases, MVPs en agility beschreven. Wij geloven dat een business gedreven aanpak de beste manier is om voortgang te maken. Zorg daarbij dat je jouw oplossingen bouwt in het kader van een totale architectuur en geen eenmalige point oplossingen. De meerwaarde van deze blauwdruk is om ervoor te zorgen dat dit overzicht bewaard blijft.

Een voorbeeld. Het betreft een zoekfunctie. Afdeling A en afdeling B willen een set van services ontwikkelen om te kunnen zoeken in data. Afdeling A is een research afdeling en vooral geïnteresseerd om te zoeken in een kennisbank. Afdeling B is op zoek naar een zoekfunctie voor eindgebruikers. Vanuit architectuur hergebruik wil je deze graag bij elkaar brengen. Wie wordt de eigenaar? Kan het met één product? Welke toekomstige eisen zullen de afdelingen mogelijk gaan stellen?

6. We hebben ook ontdekt dat deze blauwdruk handig kan zijn om eigenaarschap te definiëren. Welke afdeling is verantwoordelijk voor welke service? Wij hebben de services gegroepeerd in logisch architectuur domeinen (zie architectuur overview). Maar de realiteit zal zijn dat een organisatiestructuur

hier dwars doorheen loopt. Wij konden namelijk vanuit een greenfield denken en starten zonder ons zorgen te maken over organisatorische grenzen. Desalniettemin, de discussie kun je en moet je voeren. Daarbij gaat het niet alleen om organisatorische grenzen, maar ook de afweging welke services in een private of publieke cloud worden afgenumen, welke je als SaaS afneemt of welke je afneemt als managed service, outsourcing, etc.

7. Ten slotte wijzen we je op culturele aspecten die impact hebben op de bouw van een data-gedreven organisatie. Zelf tussen domeinen zul je culturele verschillen vinden. Wees je daarvan bewust. Soms zul je de verschillen moeten oplossen, soms wil je juist gebruik maken van de verschillen!

We hebben nu beschreven hebben hoe je dit boek kunt gebruiken. Het is ook goed om aan te geven waar dit boek niet voor bedoeld is! Het is namelijk geen kant en klare referentie architectuur. Een referentie architectuur bevat een lijst met requirements, het bevat een lijst met architectuur principes en je services moeten in meer detail beschreven zijn. Het is ook geen beschrijving van allemaal raamwerken. De raamwerken waar we naar refereren zijn uitgebreid beschreven in specifieke documentatie.

De raamwerken die we hier bedoelen zijn DevOps, ModelOps, AIOps, DMBoK en de ‘garage’ methode. Probeer deze objectief te beoordelen en bepaal wat voor jouw (organisatie) relevant is. Ons doel is om ze benoemen, je bewust te maken van dat raamwerk zodat jij vervolgens kan bepalen of deze interessant is voor je. Je moet op de hoogte zijn van hun bestaan. Ze zijn enorm belangrijk om bijvoorbeeld goed samen te kunnen werken en culturele aspecten te onderkennen.

3. ARCHITECTUUR CONTEXT

Geen AI zonder IA

3.1. GEEN ARTIFICIAL INTELLIGENCE ZONDER INFORMATIE ARCHITECTUUR

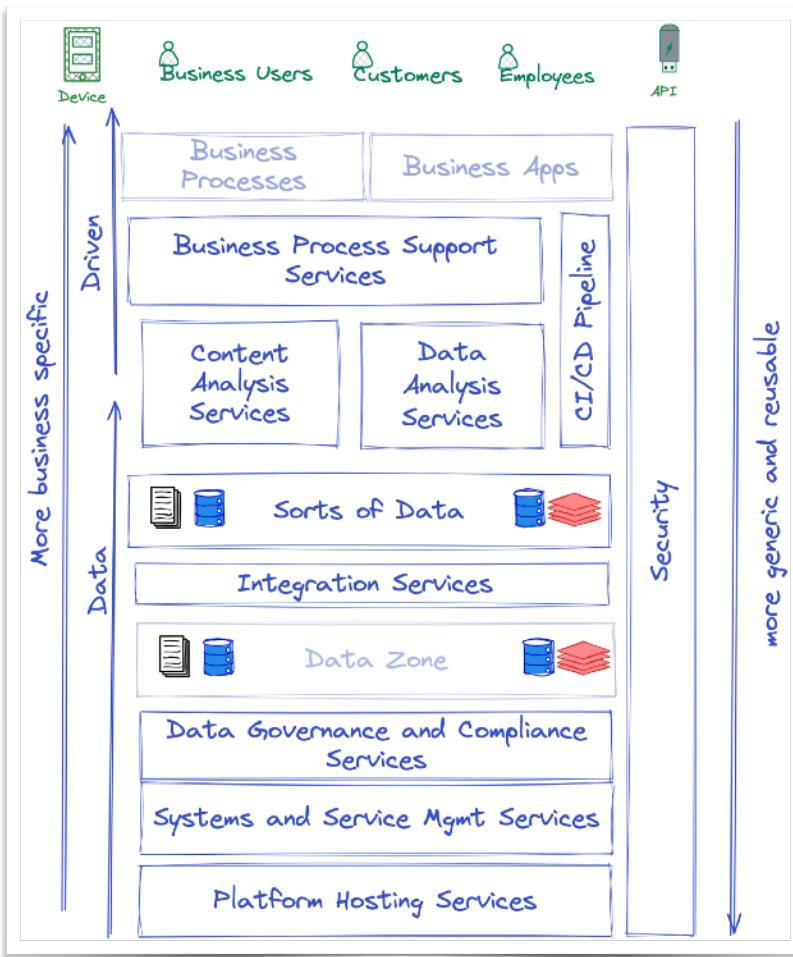
Bedrijven zien grote mogelijkheden voor het gebruik van AI (artificial intelligence). Denk hierbij aan het optimaliseren van bedrijfsprocessen en het versnellen van werkzaamheden. Onze stelling is dat je niet ver gaat komen zonder een gedegen en **holistische** (enterprise) **informatie architectuur (IA)**. Zonder deze IA kun je geen goede informatie huishouding inrichten. In dat geval blijft AI een leuk speeltje dat door een paar whizzkids wordt ontwikkeld maar nooit het licht zal zien in een productieomgeving. Waarom deze stellige mening? Vaak zien we dat informatie architectuur wel alle aspecten van data en informatie beschrijft maar de ‘umwelt’ hierbij vergeet. Wat zijn de beveiligingsaspecten van data en informatie, op welk platform gaat het draaien, hoe dragen we het over aan een productieteam? Daarnaast moet de informatie architectuur ook alle data governance aspecten bevatten of schuiven we deze verantwoordelijk plotseling door naar een ander team die het mag behandelen als een beleidskwestie?

Informatie architectuur (data architectuur) definitie

Volgens TOGAF: Een beschrijving van de structuur en interactie van de belangrijkste typen en bronnen van gegevens van de onderneming, logische gegevensmiddelen, fysieke gegevensmiddelen en bronnen voor gegevensbeheer.

3.2. ARCHITECTUUR OVERZIJKT VOOR EEN DATA-GEDREVEN ORGANISATIE

In ons architectuur overzicht wordt een set gemeenschappelijke diensten (services) gedefinieerd die zeer nuttig zijn voor een data-gedreven organisatie. We hebben deze ondergebracht in logische architectuur domeinen. We hebben discussies gehad of we het



Architectuur overzicht

model van onder naar boven moesten beschrijven (wat de meest logische benadering lijkt vanuit het standpunt van een architect) of van boven naar beneden? We kozen voor de tweede benadering (van boven naar beneden) om vanuit een business perspectief te redeneren, maar er is naar onze mening geen goede of slechte keuze.

1. **Business processen.** De bedrijfsprocessen zijn de primaire (en secundaire) functies van een organisatie. Elke organisatie heeft deze functies om de bedrijfsvoering uit te voeren. Bedrijfsprocessen verschillen per branche. Het kunnen de resultaten zijn van verkooptransacties die de afgelopen maand zijn uitgevoerd. Het kan een lijst zijn met taken die die dag moeten worden uitgevoerd. Omdat het om specifieke bedrijfsgerelateerde diensten gaan, zullen we hier niet uitgebreid op ingaan.
2. **Business Apps.** De (core) Business Applications vormen de kernsystemen van de organisatie en ondersteunen de primaire bedrijfsfuncties. Deze systemen worden vaak de Systems of Record genoemd en bevatten de operationele bedrijfsgegevens. Dit is de databron voor de Systems of Insight. Ook deze vallen buiten de scope van onze blauwdruk.
3. **Business Support Services.** De Business Support Services zijn de middelen die gebruikers ondersteunen en de mogelijkheden geven om hun activiteiten efficiënt uit te voeren. Typische services voor een data-gedreven organisatie zijn zaken zoals rules management, case management en business process management. Rules management gebruikt data of datamodellen om input te leveren aan een bedrijfsproces of een casus om zo weloverwogen beslissingen te kunnen nemen.

4. **Content Analysis Services.** We hebben lang gediscussieerd over de vraag of we Content- en Data-analyse Services in één groep moesten combineren: vanuit een service perspectief bestaat er overlap tussen die twee. Bijvoorbeeld de anonimiserings services. We hebben echter ook diensten geïdentificeerd die alleen van toepassing zijn op het gestructureerde of ongestructureerde domein. Een document Overview Service of een Q&A-service is bijvoorbeeld alleen van toepassing op ongestructureerde gegevens. Een tweede reden is dat de technologieën voor de implementatie van deze diensten vaak sterk van elkaar verschillen ondanks de conceptuele overeenkomsten.
5. **Data Analysis Services.** Wat Content Analysis Services zijn voor ongestructureerde data, zijn Data Analysis Services voor gestructureerde data. Zowel de Content Analysis Services als de Data Analysis Services worden beschouwd als het *vooraanzicht* van een gebouw voor een data-gedreven organisatie. Het geeft inzicht in data! Van feitelijke historische gegevens tot actuele inzichten, maar ook voorspellende uitkomsten en zelfs voorschrijvende suggesties over hoe te handelen op basis van bepaalde informatie. Waarom hebben we het hier alleen over het *vooraanzicht* en niet over het hele gebouw? De reden om dit het *vooraanzicht* van een gebouw te noemen is dat Data- en Contentanalyse services niet goed kunnen functioneren als de fundering van dat gebouw niet goed is. In de praktijk zien we dat de focus op het aanzicht ligt, terwijl de complexiteit en het werk in het fundament zit: het verzamelen, begrijpen, beheren en classificeren van de data. Een typische rol die wij aan de curator hebben toegewezen. De rollen die gebruik maken van de data-analyse services zijn doorgaans analisten en data scientists.
6. **Sorts of Data.** Wat kun je doen zonder data? Niets! De Sorts of Data zijn een verzameling van bepaalde soorten data, nog

meer data en nog veel meer data in welk formaat dan ook. Niet alle data hoeven fysiek op locatie beschikbaar te zijn (zie Datazone), maar kunnen ook virtueel uit de cloud worden opgehaald.

7. **Integration Services.** Toegang tot gegevens wordt verzorgd door Integration Services. Een veel voorkomende aanpak is het gebruik van API's om bijvoorbeeld een proces en een informatie service te integreren. De reikwijdte van Integration Services is echter veel breder: het is een verzameling services die elk type integratie mogelijk maakt, van het transporteren van grote, we bedoelen echt hele grote bestanden, tot het op afstand verwerken van gegevens. Van het traditioneel kopiëren tot het toegankelijk maken van data via views.
8. **Data Zone.** Datazones zijn de locaties waar de gegevens zich bevinden. Dit kan On-premises zijn, dat is in je eigen datacenter, of data die zich op een andere locatie bevinden, b.v. datacenter van een ander bedrijf of gegevens opgeslagen in de cloud.
9. **Data Governance & Compliance Services.** Data governance helpt je te begrijpen welke gegevens je hebt, waar die gegevens zich bevinden en hoe deze mogen worden gebruikt, (idealiter via geautomatiseerde zelfbedieningsmechanismen). Met compliance kunnen we aantonen dat we voldoen aan wet- en regelgeving. Deze diensten bieden oplossingen en tools voor data inzicht, datagebruik en datakwaliteit. Het biedt ook een fundament dat de samenwerking tussen business en IT ondersteunt en informatie integratie mogelijk maakt.
10. **Systems and Service Management Services.** De hybride cloud vereist services voor het beheer van zowel de operationele aspecten als de diensten aspecten. Hoewel de focus ligt op het beheer van containers, wat wordt uitgevoerd

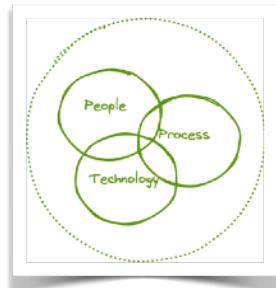
door Kubernetes, moet het Systems and Service Management de hele keten overzien en dus ook rekening houden met het bestaande enterprise platform: de Systems of Record.

11. **Platform Hosting Services.** De data-gedreven organisatie werkt het liefst met cloud concepten. Hoogstwaarschijnlijk zal dit een private cloud omvatten met mogelijkheden om dit uit te breiden naar de publieke cloud. Het moet hardware-agnostisch zijn met duurzame technologie. Denk bijvoorbeeld aan het optimaal inzetten van dure GPU-servers voor het uitvoeren van AI-modellen.
12. **CI/CD Pipeline.** De CI/CD Pipeline is het vehikel om (micro)services te produceren. We hebben tools nodig om onze code en configuraties op te slaan, om onze code te ontwikkelen, om de omgeving te bouwen, om deze te testen, en uiteindelijk moeten we deze na een succesvolle test in productie nemen. Al deze diensten maken deel uit van de CI/CD Pipeline.
13. **Data Security Services.** Wij hopen dat we nog steeds je aandacht hebben. We voelden ons een beetje ongemakkelijk om Data Security Services als laatste domein te beschrijven. Nu je een overzicht hebt van de data-gedreven architectuur, moet je de relevantie van de beveiliging wel voelen. Een deel van de gegevens die je gebruikt, zijn openbare gegevens, een deel daarvan zijn gevoelige gegevens. Wat voor soort gegevens zijn het en wie heeft er toegang toe? Zeker een apart hoofdstuk waard! Houd er rekening mee dat de nadruk hier ligt op **DATA**. Wij houden ons in deze context niet bezig met infrastructuur beveiliging zoals firewalls.

3.3. FRAMEWORKS (RAAMWERKEN)

Met al deze services beschikken we over de technologie om een data-gedreven organisatie te runnen. Technologie is een belangrijk stuk om de puzzel op te lossen, maar zonder de juiste bekwame mensen en een passende cultuur zijn we niet in staat om data om te zetten in waardevolle inzichten. De IT-industrie heeft diverse raamwerken ontwikkeld om elk aspect van de drie dimensies Mens, Proces en Technologie te adresseren. De volgende raamwerken zijn bruikbaar bij het ontwikkelen van een data-gedreven organisatie:

- Garage. De methode om de behoeften van de verschillende belanghebbenden te begrijpen en samen met gebruikers te werken aan een minimal viable product (MVP).
- DevOps. Er wordt gebruik gemaakt van een DevOps procesaanpak om bedrijfsprocessen te realiseren door applicaties te bouwen die (door de Garage) geïdentificeerd zijn.
- ModelOps. De bedrijfsprocessen zullen inzicht in data nodig hebben. Dit kan gebaseerd zijn op AI-modellen. ModelOps is een proces om datamodels te ontwikkelen, te trainen en deze in een productieomgeving in te zetten en te monitoren.
- AIOps. De gehele operationele omgeving moet holistisch worden beheerd. Dit omvat containeromgevingen maar ook andere bestaande platforms. AI wordt steeds vaker ingezet om dit beheer te ondersteunen.
- DMBok. Een belangrijke standaard voor het inrichten van een goede structuur voor data is het DMBok raamwerk. DMBok (Data Management Body of Knowledge) dient als een raamwerk voor het begrijpen en implementeren van effectieve datamanagement praktijken.



People, process,
technology

Verderop in het boek zullen we meer informatie geven over deze raamwerken.

34. PEOPLE

Een van de complexiteiten van een data-gedreven organisatie is de ervaring, kennis en vaardigheden die specialisten op dit vlak nodig hebben. Let op! Deze kennis komt niet aanwaaien.

Aan de ene kant hebben we de wereld van ongestructureerde data, hiervoor is doorgaans de expertise nodig van een bibliothecaris (tegenwoordig Curator). Dit is volgens ons één van de oudste beroepen ter wereld op het gebied van informatie huishouding. Denk bijvoorbeeld aan de bibliothecarissen van Alexandrië die alle papyrusrollen van deze beroemde oude bibliotheek catalogiseerden en classificeerden!

Het organiseren van ongestructureerde data door deze data te classificeren met behulp van metadata en het toepassen van ontologieën en taxonomieën is complex. De juiste keuzes zijn belangrijk om het classificeren makkelijker te maken, maar ook om de gegevens (ja, de papyrusrollen) makkelijker te vinden.

De aanpak die de bibliothecarissen uit Alexandrië hanteerden is van alle tijden: het scheiden van inhoud en metadata (index). Ons huidige bibliotheek systeem is nog steeds gebaseerd op dit principe (maar ook onze huidige IT-systeem!). Wanneer je een boek, film of muziek huurt, zoek je in de index en dat verwijst naar de positie in de bibliotheek.

De bibliotheek is onderverdeeld in categorieën en de media zijn dienovereenkomstig georganiseerd. Je begrijpt dat het wijzigen van categorieën of het herstructureren van een bibliotheek een enorme hoeveelheid werk betekent.

Aan de andere kant van dit spectrum hebben we de gestructureerde data. De wereld van onder andere de data-scientist. De waarde van gestructureerde data in de context van

de data scientist zit in de correlatie van grote hoeveelheden data, niet in de individuele records.

Het toepassen van AI-technieken en -modellen en het uitvoeren daarvan op grote hoeveelheden data bevindt zich helemaal aan de andere kant van het spectrum.

Daarom vinden wij het belangrijk om onderscheid te maken tussen gestructureerde en ongestructureerde data: Verschillende beroepen, verschillende omgevingen met eigen tools leiden tot verschillende werkprocessen. In tegenstelling tot wat je soms in product artikelen leest, wordt naar onze persoonlijke overtuiging deze verschillende technieken niet samengevoegd in één oplossing. De kenmerken zijn te verschillend.

3.5. KAN AI DE KWALITEIT VAN BEDRIJFSPROCESSEN VERHOPEN?

Kunstmatige intelligentie is een belangrijke technologie in een veerkrachtige data-gedreven organisatie. Wanneer je een data-gedreven organisatie wordt, is het niet langer haalbaar om data handmatig te verwerken en te interpreteren. Dat geldt zowel voor gestructureerde als voor ongestructureerde data.

Een gecontroleerde toepassing van AI zal mogelijkheden bieden om de kwaliteit van de besluitvorming te verbeteren zolang de menselijke factor betrokken blijft en beslissingen kunnen worden herleid.

Dit is geen gemakkelijke opgave; het vereist een goed fundament dat inzicht geeft in hoe een algoritme zich gedraagt, welke trainingsgegevens zijn gebruikt, welke operationele gegevens zijn gebruikt enz.

Met andere woorden, het ondersteunt de mens met de grondgedachten achter de data-gedreven beslissingen, maar biedt de mens (gelukkig) de mogelijkheid om deze beslissingen terzijde te schuiven.

3.6. HYBRID CLOUD

Deze blauwdruk voor een referentie architectuur is product agnostisch. Tijdens de vervolgstappen moet je echter keuzes maken bij het selecteren van je producten. Wij zijn van mening dat een veerkrachtige data-gedreven organisatie cloud-onafhankelijk moet zijn. Voorkom dus lock-in door specifieke cloud producten te gebruiken.

Meestal kost een oplossing die onafhankelijk is van een specifieke cloud iets meer moeite vergeleken met het gebruik van een kant-en-klaar product van de cloud provider. Het maakt je positie in een multicloud strategie sterker doordat je cloud agnostisch blijft.

Het heeft nog een aantal voordelen:

- Je bent niet afhankelijk van een public cloud provider; het werkt ook in je eigen private cloud.
- Het dwingt je om meer generieke oplossingen te gebruiken, wat leidt tot een “basis” gebruik van het product. Hierdoor worden risico’s op technische schuld vermeden.
- Het AI-gedeelte van een data-gedreven organisatie heeft een grote honger naar data. De kosten voor egress (gegevens ophalen uit de cloud) kunnen aanzienlijk zijn. Ingress (gegevens naar de cloud brengen) is doorgaans gratis. Je moet de gegevens plaatsen waar deze optimaal functioneren bij de toepassing en niet waar de cloud provider wil dat je deze neerzet.



Het is trouwens verstandig om aan te dringen op een expliciete exitstrategie, niet omdat het moet, maar omdat je dat wilt!

Er zijn drie dingen die we graag delen omdat deze nuttig kunnen zijn als best practice wanneer je in een hybride cloud omgeving werkt:

1. De eerste is Cloud Service Management and Operations (CSMO)⁷. Dit is een toevoeging op ITIL om incidenten af te handelen in een complexe multi-cloud omgeving.
2. Ten tweede zijn er activiteiten die voor een platform beheer moeten worden uitgevoerd. Specifiek de zogenaamde day-2-activiteiten. Raadpleeg een GitHub-repository voor waardevolle informatie hierover: CASE/OCP-Day2-operations⁸.
3. En ten derde is er een gratis boek: ‘The Cloud Adoption Playbook’.⁹ Het is al wat ouder, maar bevat nog steeds een uitgebreid overzicht van onderwerpen gerelateerd aan cloud adoptie. Je zou dit als een aanvulling op dit boek kunnen beschouwen, omdat het alle cloud gerelateerde onderwerpen uitdiept die we hier alleen noemen, maar niet in detail behandelen.

3.7. DATA REPATRIËRING ALS EEN SERIEUZE WAARSCHUWING

Het repatriëren van gegevens is een trend in de markt waarbij organisaties terugkomen op hun “cloud journey” naar de publieke cloud en hun data weer gaan verplaatsen naar hun lokale infrastructuur. Het traject naar de publieke cloud werd een teleurstelling en vooral de kosten van de cloud zijn de reden dat organisaties terugkeren naar activiteiten op locatie. De oorzaak dat ze terugkwamen lag vooral bij de ‘Systems of Insight’.

Hoe komt dat?

In de beginjaren van de cloud was een van de grote beloften ‘cloud is goedkoper’! Dat argument werd, op zijn zachtst gezegd, discutabel. In 2008 daalde het gebruik van IT door de financiële crisis en realiseerden bedrijven zich dat ze hun IT-kosten niet

⁷ <https://www.ibm.com/cloud/architecture/content/course/csmo-advocate/>

⁸ <https://github.ibm.com/CASE/OCP-Day2-operations>

⁹ <https://www.ibm.com/cloud/architecture/adoption/the-cloud-adoption-playbook/>

konden terugschroeven. Daarnaast was gebruiksinformatie gedimensioneerd op maximaal gebruik en dat resulteerde in aanzienlijke overcapaciteit. Het idee was dus om een flexibele IT-organisatie te creëren die zich kon aanpassen aan de variabele vraag (kosten). We gaan naar de Cloud!

Het probleem voor sommige data-intensieve bedrijven bleek in de hoge kosten van het gebruik van data in combinatie met de publieke cloud (data ophalen, lokaal verwerken en terugzetten in de cloud). Bedrijven keerden daarom terug naar IT-diensten op locatie. Echter, doordat ze de voordelen van de cloud hebben ervaren, adopteerden ze cloud technologie als basis voor hun eigen diensten. Leveranciers omarmden deze nieuwe trend en ontwikkelden private cloud-oplossingen om aan de nieuwe vraag te voldoen. Nutanix is een van de bedrijven die deze kans met succes heeft omarmd.

Minder zichtbare kostenfactoren in de publieke cloud zijn die van gegevenstransport en primaire opslag. We zeggen niet dat je jouw gegevens niet in de openbare cloud moet plaatsen, maar je moet zeer voorzichtig zijn met waar je jouw gegevens verwerkt en waar je ze opslaat. In de blauwdruk beschrijven we een Distributed Processing Service aan die de verwerking kan doen op de locatie waar de gegevens zich bevinden en alleen het resultaat retourneert. Dat zou een mogelijke oplossing kunnen zijn om dit probleem aan te pakken.

3.8. REALTIME ANALYTICS

Iets anders dat we graag willen benadrukken, is de mogelijkheid om realtime analyses uit te voeren. Realtime analyses kunnen aanzienlijke voordelen voor je bedrijf opleveren. Je bedrijf kan actie ondernemen op het moment dat de transactie plaatsvindt. Dat is data-gedreven to the max!

Dit werkt in twee stappen. De eerste stap is het creëren van een model op basis van historische gegevens. De tweede stap is om

dat model te gebruiken en het te ‘injecteren’ op het moment dat de transactie plaatsvindt. Zo is er een model ontwikkeld dat een frauduleuze creditcard transactie detecteert. In het verleden werd dat model uitgevoerd op basis van een lijst met transacties uit het verleden en ontdekte je de frauduleuze transacties nadat ze hadden plaatsgevonden. Maar nu wil je de controle uitvoeren op het moment dat de transactie plaatsvindt en heb je de mogelijkheid deze transactie te blokkeren. Dit is een voorbeeld van een optimale integratie tussen de Systems of Record en Systems of Insight.

En dan nu een lezerswaarschuwing!!

We moeten iets opschrijven over een van onze eigen producten. Vergeef ons onze enthousiasme en als je denkt dat je beter af bent met een andere oplossing voor deze real-time transactionele uitdaging, is dat uiteraard ook prima. Het gaat om de 'IBM Z' en 'LinuxONE'. Om preciezer te zijn gaat het om de 'Telum'-processor in die systemen. De processor combineert AI-mogelijkheden en CPU op één processor.

Stel je voor dat transacties door de CPU worden afgehandeld, terwijl er tegelijkertijd een parallel AI-proces is dat controleert of de transactie geldig is. Met de wetenschap dat 80% van alle financiële transacties ter wereld door dit soort servers wordt afgehandeld is de toepassing van realtime analyses voor financiële transacties zo gek nog niet. Er is nog een CPU met een soortgelijk concept, namelijk de Mac M1-processor, deze bevat GPU en CPU op dezelfde chip. Maar je zet je Mac niet als bedrijfsserver in een datacenter....

Einde van de commercial, we gaan verder met het boek!

3.9. SMART PORT: HAVENBEDRIJF ROTTERDAMMER EN INTERNET OF THINGS

Hier een voorbeeld van een mooie high tech oplossing in het havenbedrijf van Rotterdam. Het toepassingsgebied is hydro/

meteo en is gebaseerd op Internet of Things (IoT). Middels een netwerk van sensoren in het havengebied kan men accurate en actuele water- (hydro) en weergegevens (meteo) uitlezen en interpreteren voor het plannen en managen van het scheepsverkeer. Door het genereren van real time informatie kan er op het gebied van decision management snel worden gereageerd zodat de dienstverlening naar klanten van het havenbedrijf optimaal is. Met technieken als real time analytics, artificial intelligence en blockchain worden kademuren, meerpalen, waterwegen en verkeersborden gemonitord door sensoren en worden deze gegevens uitgewisseld met andere autonome systemen. Je ziet het al aankomen: nog even en de scheepvaart kan volledig zelfstandig manoeuvreren in het havengebied van Rotterdam.

Binnen het nu opgeleverde hydro/meteo-systeem worden data over waterstanden, stroming, zoutgehalte, windsnelheid, windrichting en zicht verkregen via een combinatie van 44 sensoren in de haven.

Daarmee draagt de applicatie bij aan het verminderen van wachttijden en optimalisatie van aanleg-, laad/los-, en vertrektijden. Dankzij de technologie kan bijvoorbeeld nauwkeuriger voorspeld worden wat, afhankelijk van het water niveau, het beste tijdstip is om aan te meren en te vertrekken, onder garantie van een maximale lading (diepgang).

Door gebruik te maken van wereldwijde en open standaarden is het voor externe partijen nog nooit zo eenvoudig geweest om te koppelen met deze real-time data. Per dag verwerkt het platform nu al zo'n 1,2 miljoen datapunten voor modellen, systemen en gebruikers.¹⁰

Een mooi voorbeeld van een date-gedreven organisatie!

¹⁰ <https://www.portofrotterdam.com/nl/nieuws-en-persberichten/havenbedrijf-rotterdam-neemt-internet-things-platform-gebruik>

3.10. ROADMAP EN PLANNING

Wanneer je met alle data-gedreven diensten wordt geconfronteerd, kan het behoorlijk overweldigend overkomen. Zeker als je organisatie beperkt van omvang is. Maar maak je geen zorgen, je hoeft niet alles (tegelijkertijd) te doen! De beste aanpak is om eerst jouw meest waardevolle use-cases te definiëren. Gebruik de blauwdruk als referentiearchitectuur om je use-cases in kaart te brengen. Met deze aanpak vind je de belangrijkste services. Ontwikkel deze services als eerste!

II. DE BLAUWDRUK

De blauwdruk is net een metrokaart, het geeft een mooi overzicht van alle mogelijke bestemmingen

4. STRUCTUUR VAN DE BLAUWDRUK

De blauwdruk is uitgewerkt op posterformaat (A0). Om een goede structuur in de uitleg van de verschillende domeinen van de poster te krijgen, lopen we door de poster van boven naar beneden (zie het architecturoverzicht in het hoofdstuk ‘Architectuur Context’). We hadden het net zo goed vanonder kunnen opbouwen, maar we hebben gekozen voor de top-down aanpak om ons vooral te richten op het gebruik van data in bedrijfsprocessen! Technologische aspecten zoals platform hosting diensten zijn zeker ook belangrijk, maar zijn vooral middelen om het doel te bereiken.

Elk architectuur domein wordt beschreven als een set van services of capabilities. We maken in ons boek geen onderscheid tussen services en capabilities en we gebruiken deze door elkaar heen.

Persona's in deze blauwdruk

	Business user	De business user (zakelijke gebruiker) staat centraal in de bedrijfsvoering. In een data-gedreven organisatie vertrouwt de zakelijke gebruiker op betrouwbare inzichten.
	Developer	De developer (ontwikkelaar) is actief in een aantal domeinen. Bij bedrijfsondersteunende diensten bouwt de ontwikkelaar oplossingen om de bedrijfsproductiviteit en de automatisering van bedrijfsprocessen te verbeteren.

	Data scientist	De data scientists zijn experts op het gebied van data-analyse. Ze beschikken over vaardigheden om complexe problemen op te lossen, zijn in staat bedrijfsresultaten te voorspellen en zakelijke besluitvorming te ondersteunen.
	Curator	De data curatoren zijn verantwoordelijk voor het onderhouden en beheren van metadata. Zij catalogiseren de gegevens en zorgen ervoor dat de gegevens worden gebruikt zoals ze bedoeld zijn. Ze overbruggen de kloof tussen de wereld van informatietechnologie en data science.
	Data engineer	Data engineers zijn technische specialisten die verantwoordelijk zijn voor het ontwerpen en onderhouden van de architectuur van datasystemen. Zij zijn betrokken bij de processen voor verbetering van de data kwaliteit. Ze zijn ook verantwoordelijk voor de processen van modellering, mining, verificatie en data acquisitie.
	Tester	Er zijn veel verschillende soorten testers. In iedere context toetsen zij of het product voldoet aan de gestelde functionele en niet-functionele eisen.
	Engineer	Engineers zijn verantwoordelijk voor het ontwerp, de installatie en het onderhoud van enterprise IT-systemen.

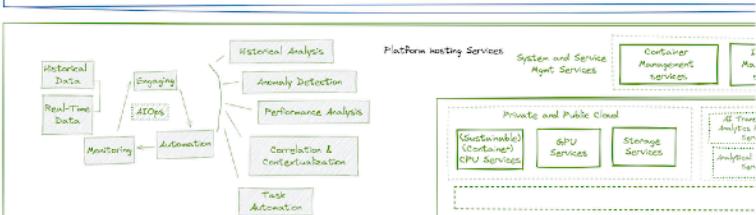
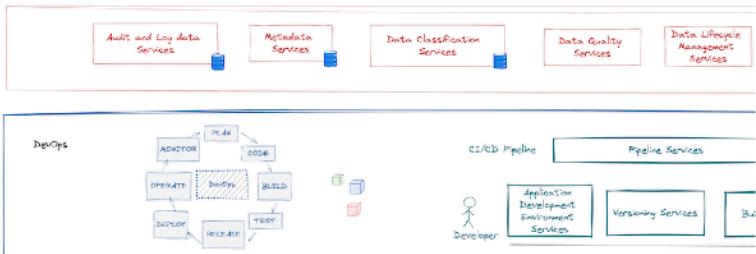
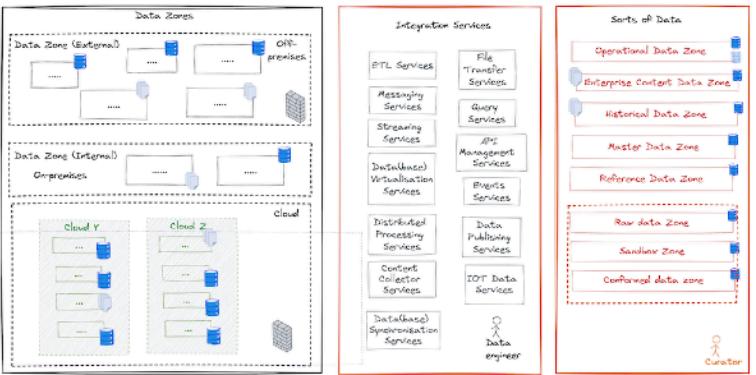
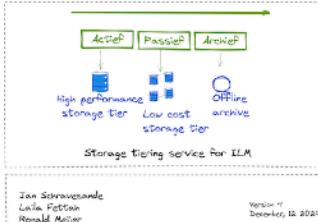
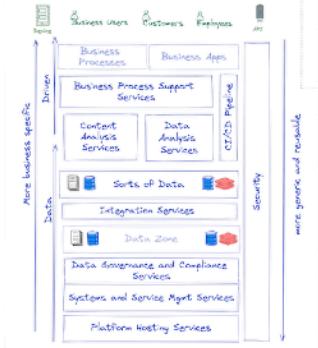
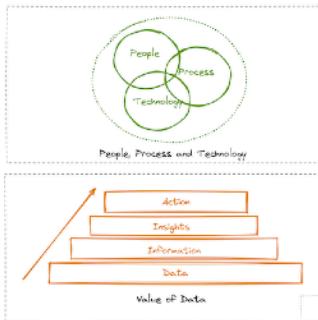
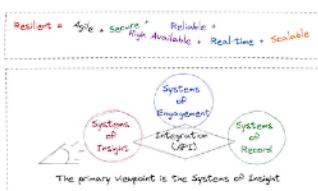


SRE team

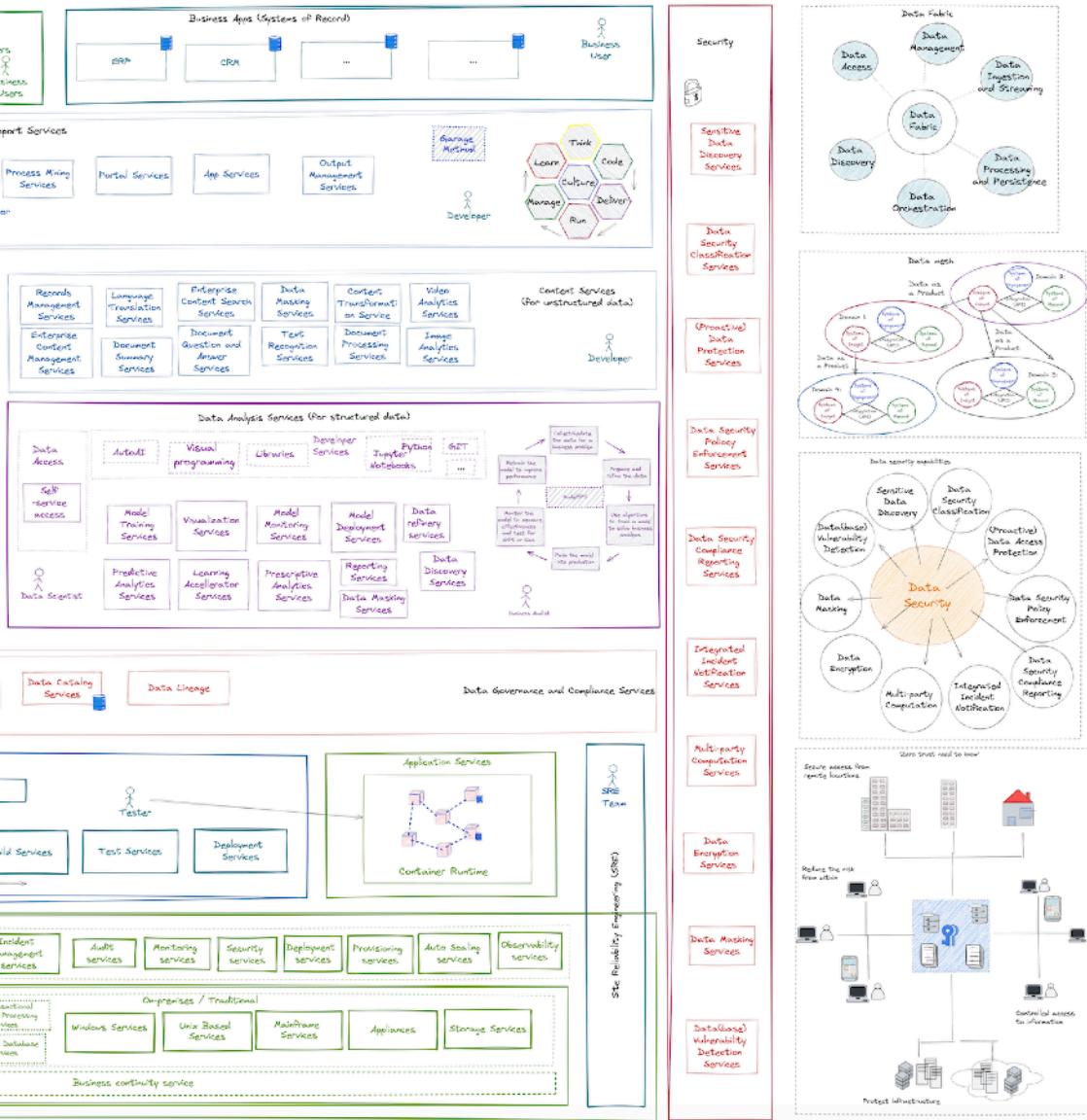
SRE-teams (Site Reliability Engineering) zijn verantwoordelijk voor het proactief bouwen en implementeren van diensten die de IT-organisatie helpen beter te worden in hun werk. De focus van een SRE ligt op de totale stack van technologie die een dienst levert. Een SRE heeft daarom kennis van infrastructuur, middleware en applicatie software.

Enterprise Architecture Blueprint

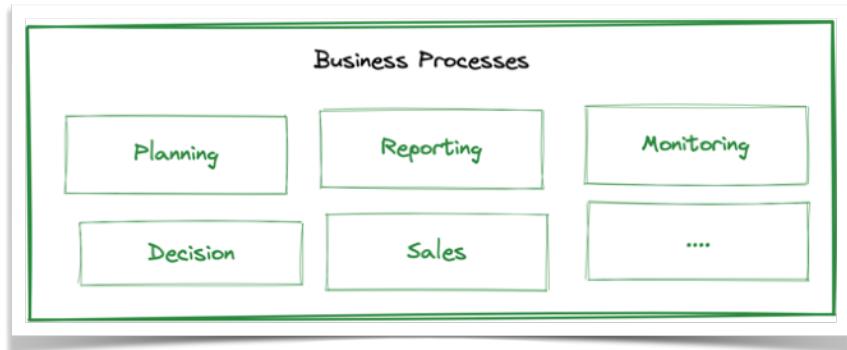
A data-driven organization focus on the system of insight to integrate data from different sources to collect insight in data and apply this in their (primary) processes. This insight could range from simple analytics till sophisticated Deep Learning models. This architecture is a collection of the services to train and deploy these models, develop apps (microservices) to deploy those and make those models available to the (primary) business processes of your company.



for a Resilient Data-Driven Organisation



5. BUSINESS PROCESSES



Business Processes

Bedrijfsprocessen zijn doorgaans gerelateerd aan een specifiek bedrijf. Hoewel er enkele algemene processen zijn die verband houden met de ondersteuning van het bedrijf, zijn de kern bedrijfsprocessen vrij uniek per branchetype.

Primaire bedrijfsprocessen zijn bijvoorbeeld:

Voor een bedrijf dat producten maakt:

- Inkoop
- Productie proces

In de detailhandel bijvoorbeeld:

- Assortimentsbeheer
- Reclame en marketing

Organisaties die reizen organiseren en transporten plannen:

- Planning
- Kaartverkoop

Naast primaire processen hebben bedrijven ook ondersteunende bedrijfsprocessen. Deze staan bekend als COPAFIJTH of SCOPAFIJTH:

- Security
- Communicatie
- Organisatie
- Personeel
- Administratie
- Financiën
- Informatie
- Juridische aspecten
- Technologie
- Huisvesting

We moeten onszelf twee belangrijke vragen stellen:

1. *Hoe kunnen organisaties hun bedrijfsprocessen voeden met kwaliteitsinformatie om meer data-gedreven te worden?*
2. *Welke informatie is nodig om een bepaald proces te voeden?*

Opdracht:

Werk je eigen datavereisten uit door gebruiksscenario's te ontwikkelen en identificeer de (data)diensten voor elk van de gebruiksscenario's (zie hoofdstuk achterin het boek 'En Verder').

Enkele ideeën met betrekking tot bedrijfsprocessen voor een data-gedreven organisatie

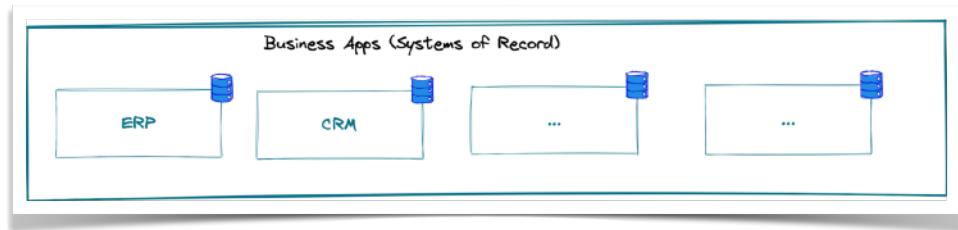
Decision Process	<p>Alle organisaties moeten beslissingen nemen. Jouw organisatie heeft bijvoorbeeld een nieuwe CTO nodig. HR moet beslissen wie de beste persoon voor de functie is. Laten we het compliance-gedeelte van de oplossing niet bespreken, maar het proces doorloopt alle HR-dossiers en leidt tot de beste persoon voor de taak.</p> <p>En raad eens: ja, jij bent het!</p>
Planning Process	<p>Bijna alle bedrijven hebben te maken met planningen. Zeker als je in de logistiek zit. Je bent bijvoorbeeld pakketbezorger en je wil weten wat de beste route is voor bezorging. Een ander voorbeeld: je bent politieagent en wil weten wat de beste surveillance route is. Planning gebeurt overal. Een goed planningsproces kan tijd, geld en zelfs levens besparen. Optimalisatie-algoritmen kunnen planningen doorrekenen en optimaliseren. Wat is jouw belangrijkste planningsproces? Wat zou je winnen als dat maar 1% beter zou zijn? Nee, denk niet alleen aan geld!</p>

Opdracht:

Wat zijn de drie belangrijkste bedrijfsprocessen in jouw organisatie en hoe kun je deze verbeteren door data te injecteren?

Reporting Process	<p>Er zal altijd behoefte blijven aan het genereren van rapporten voor het management die de status van de organisatie weerspiegelen: bijvoorbeeld maandelijkse verkopen. Om een startpunt voor een use-case te vinden kun je jezelf de vraag stellen: "Wat zijn de belangrijkste rapportages in mijn organisatie?". Kun je deze verrijken met aanvullende informatie die besluitvormers een beter inzicht kunnen geven? Je verzorgt bijvoorbeeld maandelijks een rapportage over de verkoop van ijs. Zou het zinvol zijn om informatie toe te voegen over het weer, de smaken en de concurrentie?</p>
Sales Process	<p>Alle organisaties hebben een soort verkoopproces. Of het nu een overheidsorganisatie is die paspoorten uitgeeft of een kruidenierswinkel die aardappelen verkoopt. Zij 'verkopen/leveren' iets en er zijn waarschijnlijk veel inzichten die nuttig kunnen zijn in dit primaire proces, zoals kansen, verkopen, verliezen, vooruitzichten, toekomstvoorspellingen, etc.</p>
Monitoring Process	<p>Het Monitoring Process maakt gebruik van de Business Process Monitoring Services om inzicht te krijgen in de status van interne en externe processen. Het kan bijvoorbeeld de voorraad monitoren en waarschuwt je als het aantal producten onder een bepaalde drempel komt.</p>

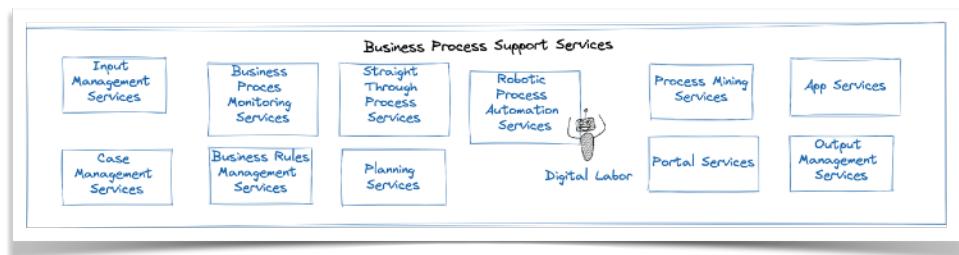
6. BUSINESS APPS (SOR)



Business Apps

De business apps vormen de kernsystemen van de organisatie en ondersteunen de primaire bedrijfsfuncties. Zoals eerder vermeld zijn deze systemen vaak het resultaat van jarenlange ontwikkeling en nog steeds onderhevig aan functionele en technische veranderingen. Daarom kan het soms een technische uitdaging zijn om de waardevolle data van deze systemen te ontsluiten. Vooral als we de (real-time) data willen gebruiken in AI-processen. Veel organisaties gebruiken ERP-software (SAP, Oracle, Microsoft) en hopelijk gebruiken ze het als COTS (commercial off the shelf), anders staan deze organisaties voor dezelfde uitdaging als degenen die hun eigen systemen hebben ontwikkeld (maatwerk). Upgraden naar een nieuwe versie van ERP is bijna of zelfs een grotere uitdaging als je te veel functionaliteit van eigen bodem in een standaard ERP bouwt. We gaan niet in detail in op zakelijke apps, omdat deze organisatie specifiek zijn. In de A0-poster kun je jouw kernsystemen in het Business Apps-domein plaatsen.

7. BUSINESS PROCESS SUPPORT SERVICES



Business Process Support Services

Business Process Support Services zijn generieke functionele bouwstenen die beschikbaar zijn om bedrijfsproces diensten te ontwikkelen. Voorbeelden zijn rule management en proces management. Deze diensten zijn niet persé bedoeld voor de Systems of Insight die we eerder bespraken, maar maken deel uit van de data-gedreven organisatie. Een data-gedreven organisatie gaat over ‘data’ en gaat over ‘gedreven’. De focus voor de Business Process Support Services ligt op het “gedreven” deel. Het maakt gebruik van data en stelt die beschikbaar aan de bedrijfsprocessen.

Case Management Services	<p>Case Management is een specifieke invulling van proces management. Het wordt gebruikt wanneer informatie moet worden verwerkt via een ongedefinieerde reeks van taken. De uitkomst van een taak bepaalt de volgende taak die moet worden uitgevoerd. Het verzamelt gaandeweg informatie en inzichten en het wordt een case (dossier). Als je bijvoorbeeld een snelheidsovertreding hebt begaan, wordt er een bedrijfsregel gehanteerd: wanneer de overtreding meer dan 50 km bedraagt, wordt de overtreding een zaak. Als het minder is, gaat het via Straight Through Processing, waardoor je een boete krijgt en misschien de link naar een foto waar en wanneer de overtreding plaatsvond. Dure foto's!</p> <p>PS Een zaak die al jaren niet meer de warme aandacht heeft, wordt een "cold case" genoemd.</p>
Straight Through Processing Services	<p>Straight Through Processing Service of STP is de andere implementatie van Business Process Management (BPM). STP verwerkt een stukje informatie via een vooraf gedefinieerde reeks taken. Stappen en volgorde binnen het bedrijfsproces zijn vooraf bepaald. Bedrijven proberen het grootste deel van hun processen via STP te laten verlopen, omdat dit geen of zeer beperkte gebruikersinteractie vereist. STP integreert met Case Management. Zodra het proces niet meer via het standaardproces kan worden afgehandeld, wordt het overgedragen aan Case Management. En ook andersom. Als een case weer een standaardproces wordt, stroomt deze door naar STP.</p>

Business Rules Management Services	<p>Business Rules Management Services zijn services die allerlei soorten bedrijfslogica en regels bevatten. Het wordt centraal beschikbaar gesteld en is via een eenvoudige interface, bij voorkeur een API, toegankelijk voor bedrijfsprocessen. De belangrijkste voordelen zijn:</p> <ol style="list-style-type: none"> 1. enige bron van waarheid en 2. de regels kunnen relatief eenvoudig worden gewijzigd, omdat ze niet hard gecodeerd zijn.
Process Mining Services	Process Mining Services meet en rapporteert over de stroom van procesactiviteiten en geeft advies om het proces te verbeteren.
Business Process Monitoring Services	Terwijl de mining dienst kijkt naar de flow van de processen, zijn de Business Process Monitoring Services gefocust op de actuele status van het proces. Denk hierbij aan achterstallige aanvragen, de lengte van een wachtrij, gemiddelde responstijden, etc.
Application Services (Apps)	Application Services bieden services die toegankelijk zijn voor alle zakelijke ondersteunende services. Apps worden bij voorkeur ontwikkeld als microservices in een cloud native omgeving (container runtime).

Robotic Process Automation Services	<p>Robotic Process Automation Services (ook wel digital labour genoemd) automatiseren een reeks herhaalbare menselijke interacties met een computer. Het is een oplossing om snel productiviteitsvoordelen te behalen en (saai?!) herhaalbare taken te automatiseren. Over het algemeen is RPA meer een korte termijn oplossing omdat het gevoelig is voor wijzigingen in bijvoorbeeld een gebruikersinterface. Daarom gaat de voorkeur uit naar een op bedrijfsprocessen gebaseerde oplossing met herbruikbare services.</p>
Input Management Services	<p>Organisaties beschikken over verschillende communicatie kanalen (telefoon, internet, mail) en verschillende media (mp3, mp4, papier, xml). Input management is de dienst om de binnenkomende informatie te verzamelen en te distribueren en indien nodig om te zetten in een digitaal formaat zodat deze in een IT-systeem kan worden opgeslagen. Deze gegevens kunnen vervolgens worden ingevoerd in de bedrijfsprocessen. Input Management brengt zoveel mogelijk structuur aan rond ongestructureerde informatie en slaat veelal de originele input op als uitgangspunt voor een audit trail. Input management maakt gebruik van de capabilities van Content Services. Bijvoorbeeld een tekst-herkenning service of een automatische classificatie service. Een goede Input Management Service levert voordelen op in al je processen.</p>

Planning Services	<p>Planning optimalisatie en analyse services worden gebruikt voor besluitvorming. Het is bijna een wereld op zichzelf. Om helemaal eerlijk te zijn, het is voor sommigen van ons een beetje te complex. (We zullen niet bekendmaken voor wie, omdat we denken dat dit gevolgen kan hebben voor de rest van onze carrière en we het veel te leuk vinden wat we doen (-:).</p> <p>In het verleden hadden we planningsoptimalisatie en analyse als twee afzonderlijke onderwerpen, maar deze onderwerpen zijn steeds meer met elkaar geïntegreerd. Planning analytics geeft inzicht in de actuele status en forecast.</p> <p>Planningsoptimalisatie is meer gericht op het prescriptive deel, om jouw planning te verbeteren. Het is ook een domein waar nog veel onderzoek plaatsvindt. En daar komt het: het zou het domein kunnen zijn waar Quantum computing een rol zou kunnen spelen.</p>
Portal Services	<p>Portal Services behoren primair tot de Systems of Engagement. Wij zien dashboards en formulieren als een soort portal. Dashboards zijn een manier om data te presenteren en zijn daarom meer verwant aan de Systems of Insight. Het maakt gebruik van de views in de virtualisatie service (zie Data Analysis Services). Door de Input Management Services worden de formulieren gebruikt uit de portal om op een meer gestructureerde manier informatie te verzamelen.</p>

Output Management Services	Output management neemt het resultaat van een bedrijfsproces en vertaalt dat naar een consumenten友好的 output formaat. In correspondentie met klanten kan dit een PDF, een bericht of een e-mail zijn. De uitvoer die met de gebruiker wordt gedeeld, wordt opgeslagen in een repository als onderdeel van audit informatie.
----------------------------	--

7.1. BUSINESS PROCESS MANAGEMENT (BPM) EN BUSINESS PROCESS MONITORING (BPM)

Je hoort de termen vaak door elkaar gebruikt worden maar er is een essentieel verschil! BP monitoring analyseert de operationele processen op performance en bottlenecks. BP management gaat over het ontwerpen, modelleren en optimaliseren van processen voor het verbeteren van de efficiency en de effecten. Eigenlijk is het monitoring gedeelte een component van het bredere BP management process.

Binnen het BP management zijn zes belangrijke gebieden gedefinieerd:

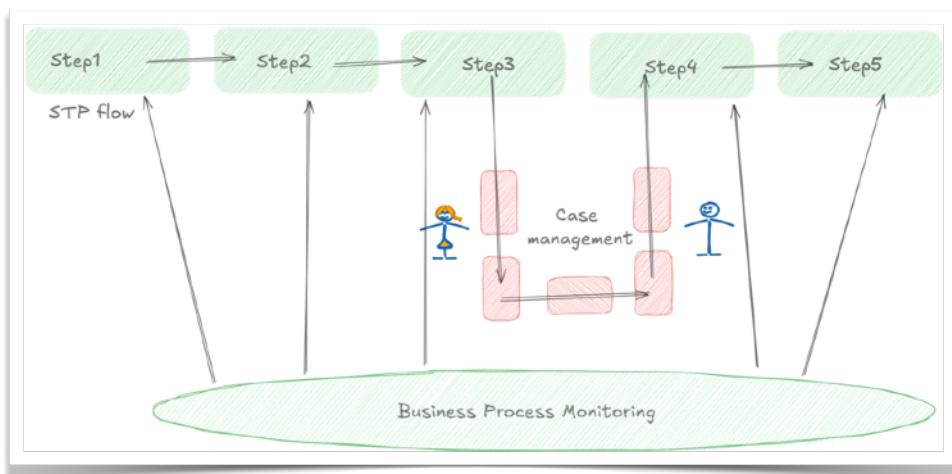
- Process modeling
- Process execution
- Process monitoring en optimalisatie
- Process governance
- Process documentation
- Process analytics

En dan heb je ook de term process mining. Wat is dan het verschil tussen process mining en process analytics? Beide technieken geven inzicht in bedrijfsgegevens rond processen maar er is wel een verschil in focus. Process mining kijkt naar business processen op basis van event logs en andere proces data. Hierbij krijg je inzicht in hoe processen in de praktijk werken en kun je dit vergelijken met het theoretische ontwerp. Process mining kan

inefficiënties, bottlenecks en andere elementen identificeren die de performance beïnvloeden van het proces, en geeft ook inzicht geven in mogelijke verbeter acties.

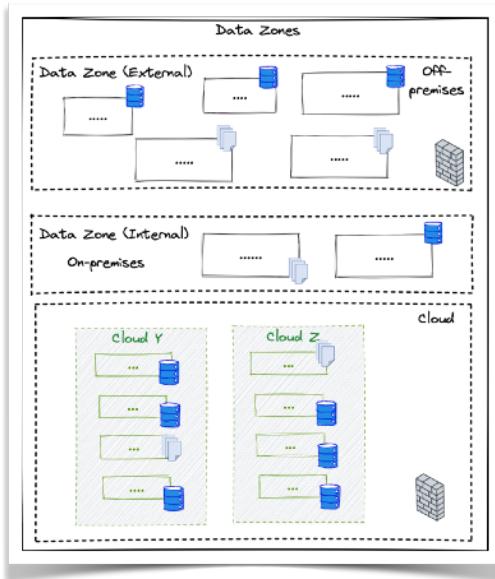
Process analytics is een breder begrip en gebruikt data en statistische methoden om meer inzicht te krijgen in business gerelateerde zaken. Analytics levert data points, KPI's en tijdsgerelateerde meetpunten: dagelijks, wekelijks, maandelijks, etc.

Dit geeft bijvoorbeeld inzichten, van klantgedrag tot financiële data. Analytics wordt toegepast op een grote variëteit aan data om metrics, KPI's en trend analyses te produceren, process mining verbind deze data points aan de taak of activiteit binnen het proces. Process mining maakt gebruik van event logs en andere process data en combineert deze andere data attributen om inzicht verschaffen in hoe processen werken in de echte wereld. Met andere woorden analytics kan je vertellen dat een process goed performt (of niet), process mining kan je vertellen waarom dat is!



Proces flow

8. DATA ZONES



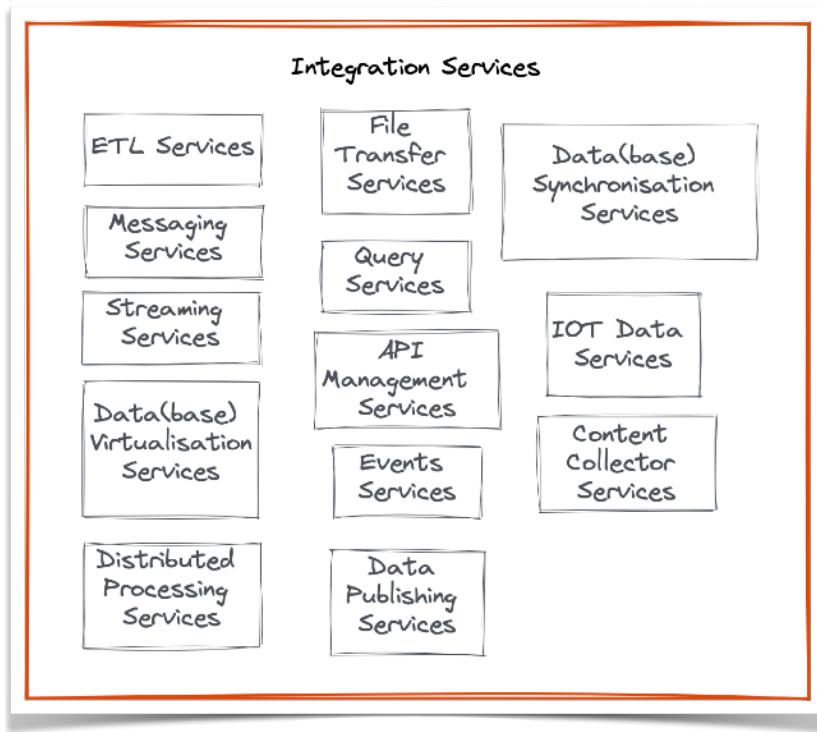
Data Zones

Data zones zijn de locaties waar de gegevens zich bevinden. Dit kan 'On-premises' zijn in je eigen datacenter of op een andere locatie, b.v. een datacenter van een ander bedrijf of in de cloud.

Uiteraard zijn alle gegevens, ongeacht de locatie, beveiligd. Wie toegang heeft tot welke data in welke context, is altijd een zeer belangrijke overweging in je architectuur.

	<p>Met externe datazones definiëren we de locatie van gegevens die geen deel uitmaken van jouw bedrijfsinfrastructuur. Het kan je eigen data zijn of van anderen. Kenmerken van externe data zones waarmee rekening moet worden gehouden: latency, beveiliging, nauwkeurigheid, transportprotocollen en betekenis.</p>
Data Zone (Internal)	<p>De interne datazone bevatten veelal je (kritische) bedrijfsgegevens.</p>
Cloud	<p>We onderkennen zowel een private als een public cloud. Typische overwegingen die je hebt bij het opslaan van gegevens in de cloud zijn de vertrouwelijkheid niveaus. Deze zijn van invloed op de vereiste versleuteling (encryptie). Versleuteling van ‘data at rest’ en ‘data in motion’ zijn algemeen voorkomende technieken. Maar als gegevens tijdens de verwerking versleuteld moeten blijven, kun je homomorphic encryptie overwegen. (Homomorphic encryptie maakt het mogelijk complexe wiskundige bewerkingen uit te voeren op gecodeerde gegevens zonder dat de data ge-decrypt hoeft te worden.) Een tweede overweging zijn de kosten van egress (uitgaande data). Bij uitgaand data verkeer haal je gegevens uit de cloud. Hoewel dit bij de Systems of Engagement of Systems of Record meestal geen probleem is, kunnen de hoeveelheden gegevens voor de Systems of Insight groot zijn. Daarom moet je overwegen waar je de verwerking uitvoert en waar je de gegevens opslaat. Distributed Processing Services (zie Integration Services) bieden een service om gegevens op afstand te bewerken.</p>

9. INTEGRATION SERVICES



Integration Services

Integration Services betreft een breed scala aan diensten om toegang te krijgen tot gegevens.

Integratie kan een enkele API-call zijn, maar ook een complex content delivery network (CDN) dat streamingdiensten levert. Een data-gedreven organisatie moet uiteindelijk een brede reeks integratie- en transportservices bieden om gegevens uit diverse

bronnen te kunnen halen. Deze gegevens kunnen door bedrijfsprocessen worden gebruikt.

In een data-gedreven organisatie kunnen een breed scala aan integratiediensten nodig zijn. In jouw roadmap bepaal je welke integratiediensten het belangrijkst zijn en begin daar als eerste mee.

Integratiediensten zijn meestal al in gebruik bij SoR en SoE. In veel gevallen kun je deze hergebruiken. De Sol hebben echter ook hun eigen karakteristieken. Zo is actualiteit vaak minder relevant, maar de hoeveelheden data kunnen enorm oplopen. De neiging bestaat dat de Enterprise Architecten tegen je zullen zeggen dat je de bestaande integratieservices moet gebruiken. Maar het kan zijn dat je dan bezig bent om een vierkantje door een rondje te persen. Dit vraagt om een architecturbeslissing. Verderop in dit hoofdstuk hebben we daarom een beslisboom opgenomen die je kan helpen om de afweging te maken bij de keuze van de integratieservice.

Opdracht:

Ga na welke type data je moet transporteren en welke integratie services hier het best bij passen. Maak je eigen beslisboom.

API Management Services	API Management Services publiceert en beheert je API's. API's zijn een manier van losse koppelingen (loose coupling) tussen applicaties. Meestal gebruiken API's het REST-protocol. De API-gateway is een van de belangrijke componenten om de beveiliging af te dwingen en de doorstroming te reguleren. API Management verzamelt gebruiksstatistieken, wat waardevolle input kan zijn voor de data scientist om inzicht te krijgen in het gedrag en de interesses van gebruikers.
Messaging Services	Asynchroon voor betrouwbaar transport van berichten en synchroon voor betrouwbaar transport met tijdsafhankelijkheid.
ETL Services	Extraheren, transformeren en laden van (grote hoeveelheden) gestructureerde gegevens van en naar (relationele) databases.
Data(base) Virtualisation Services	Gegevens zijn virtueel toegankelijk via de Database Virtualisation Services. Het is niet nodig om de gegevens naar een lokale repository te kopiëren. Het voordeel is dat de gegevens altijd actueel zijn en dat er geen lokale bronnen nodig zijn om gegevens op te slaan en te synchroniseren.
Query Services	Voert een (externe) query uit op een database om gegevens op te halen voor verwerking. SQL of Structured Query Language is een voorbeeld van een querytaal. LDAP is een andere taal waarmee je wellicht bekend bent, maar mogelijk ken je nog veel meer talen.

File Transfer Services	File Transfer Services transporteert bestanden op een gecontroleerde manier van de ene locatie naar de andere. Er zijn veel tools beschikbaar en afhankelijk van de hoeveelheid gegevens en het beveiligingsniveau kun je een keuze maken.
Data Publishing Services	Data Publishing Services maakt, veelal analytische, gegevens beschikbaar waarop anderen zich kunnen abonneren. (Publish en Subscribe)
Distributed Processing Services	Distributed Processing Services voert (data)verwerking uit bij de bron. In dat geval is het niet nodig om alle brongegevens te transporteren; alleen de resultaten worden getransporteerd. Dit vermindert egress (uitgaande data kosten) of kan nuttig zijn als een analyse moet worden uitgevoerd op actuele gegevens, in bijvoorbeeld de operationele systemen. Er zijn slechts een beperkt aantal producten beschikbaar die dit kunnen doen zonder noemenswaardige impact op de operationele systemen.
Events Services	Event Services of Event en Trigger Services detecteren gebeurtenissen en kunnen op deze gebeurtenissen reageren door een proces, een activiteit of een taak te activeren.

Data(base) Synchronisation Services	<p>Data(base)-synchronisatie houdt twee of meer dataverzamelingen identiek. Het doel is om altijd een kopie van jouw gegevens te hebben voor het geval een van de gegevensbronnen beschadigd raakt. De actualiteit van de kopie wordt uitgedrukt in RPO. Dit staat voor Recovery Point Objective. RPO=0 betekent geen gegevensverlies bij een incident. RPO=1 (uur) betekent dat je maximaal één uur aan gegevens verliest. Een voorbeeld van synchronisatie van twee databases (master - slave) geschieht door de wijzigingen die in de master plaatsvinden door te voeren op de slave. Dit gebeurt op basis van de journaalgegevens uit de database. Het vastleggen van wijzigingsgegevens heeft geen invloed op de prestaties van de database omdat deze het logbestand (journaal) gebruikt. Het is bijna realtime (RPO bijna 0). Real-time synchronisatie (RPO=0) is alleen mogelijk als de netwerk latency dit toelaat.</p>
Content Collector Services	<p>Content Collector Services brengen ongestructureerde gegevens onder controle van een Enterprise Content Management-systeem (ECM). E-mailgegevens worden bijvoorbeeld gekopieerd naar de ECM-database. Vanaf dat moment wordt de e-mail beheerd door het ECM-systeem, de wijzigingen in de e-maildatabase worden doorgegeven. Dit is een manier om je e-mailsysteem onder het beheer van de informatie levenscyclus te brengen wat een wettelijke vereiste zou kunnen zijn: b.v. wie was op welk moment betrokken bij het nemen van de beslissing?</p>

IoT Data Services

IoT Data Service is een dienst die de gegevens van edge-apparaten beschikbaar maakt. Dit kan variëren van temperatuur-sensor gegevens tot drone video's. IoT-data zijn typische data die (instant) beslissingen aansturen. De hoeveelheden zijn groot, de snelheid hoog, de relevantie kort. In het voorbeeld van F1-racen worden veel sensoren gebruikt om 1200 datapunten/sec te genereren. In de vliegtuigindustrie verzendt elk vliegtuig dat bij de gate landt 1 TB aan vluchtgegevens. Hoe zit het met onze toekomstige auto's waar we achter het stuur zitten (maar niet meer zelf besturen)?

9.1. TRANSPORT- OF TOEGANGSTECHNOLOGIE SELECTEREN

Afhankelijk van de verschillende kenmerken van de gegevens die je wilt inzien, transporteren of verstrekken, kunt je kiezen uit een breed scala aan technologieën om die taak voor je uit te voeren.

De beslissingstabel aan het einde van dit hoofdstuk is een voorbeeld om je te ondersteunen bij de keuze voor een technologie.



Data direction decision

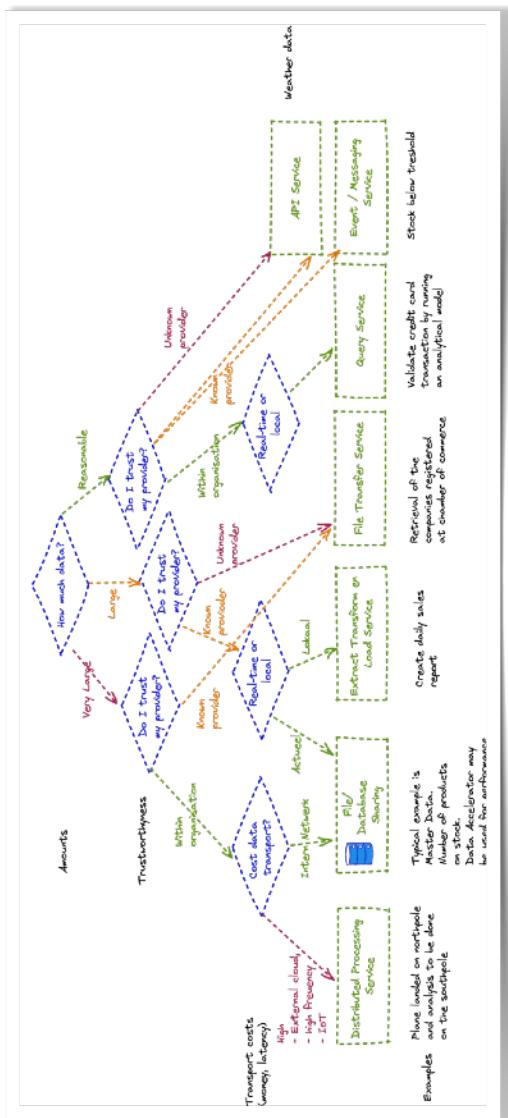
Zowel voor het ophalen van gegevens als voor het verstrekken van gegevens gelden verschillende kenmerken.

9.2. HET OPHALEN VAN GEGEVENS

Voor het ophalen van gegevens kunt je bijvoorbeeld dit script volgen:

- Hoe betrouwbaar is mijn provider?
 - Je vertrouwt hem omdat het in je eigen organisatie is
 - Je kent hem en je hebt een overeenkomst afgesloten (authenticatie/autorisatie)
 - Je kent de aanbieder niet. Het is bijvoorbeeld open data.
- Hoeveel gegevens verwacht ik op te halen?
 - Zeer groot: > 100 GByte
 - Groot: > 100 MByte
 - Redelijk: Max 100 MByte
- Hoe duur is dit transport?
 - Het kan duur zijn, omdat je een ‘boete’ betaalt voor latency of vanuit een financieel perspectief: uitgaande kosten voor een cloudprovider.
 - Het kan bijna niets zijn, want het is op je eigen terrein. Je betaalt waarschijnlijk kosten voor het verlenen van toegang ertoe en een deel van de kosten voor de routers en switches.

Hierbij geven we een voorbeeld van een beslisboom die deel zou moeten uitmaken van je oplossingsarchitectuur in de vorm van een architectuur beslissing.



Beslisboom voor het ophalen van data

9.3. HET VERSTREKKEN VAN GEGEVENS

Wanneer je gegevens verstrekt, is de eerste vraag: "waarom?"

Meteen gevuld door: "aan wie?". Er kunnen verschillende redenen zijn om gegevens aan anderen te verstrekken:

1. Het kan onderdeel zijn van de bedrijfsstrategie. Het bedrijf verstrekkt bijvoorbeeld gratis weergegevens aan het publiek, maar als bedrijven die gegevens gebruiken, moeten ze de gegevens kopen. Het bedrijf geeft een lijst met beschikbare vliegtuigstoelen op, in de hoop dat anderen die lijst zullen gebruiken om die stoelen bezet te krijgen. Je moet een evenwicht vinden tussen de kosten van het beschikbaar stellen van gegevens en de voordelen die het met zich meebrengt.
2. Je hebt een verplichting, je wilt transparant zijn of gewoon voor de lol. Als je een overheid, financiële instelling of non-profitorganisatie bent, beschik je mogelijk over interessante gegevens die je wil teruggeven aan de maatschappij of je wilt deze publiceren om je transparantie aan te tonen.
3. Data wordt door jouw bedrijfsketen overkoepelend ingezet. Als je deel uitmaakt van een supply chain, wil je jouw voorraadgegevens beschikbaar stellen aan anderen zodat zij de productie kunnen plannen.

De tweede vraag is *wie* het gebruikt. Er zijn twee aspecten aan deze vraag:

1. Is er een voordeel verbonden aan het delen van de gegevens?
2. Welke gegevens mogen ze zien?

Wij laten het aan jou over om je eigen beslisboom voor de datavoorziening op te bouwen. Maar niet zonder wat extra advies!

Het idee van een data mesh (we bespreken data mesh verderop) kan nuttig zijn.

Ook al maak je je waarschijnlijk niet zozeer zorgen over de locatie van gegevens vanuit een netwerk latency perspectief, als je de

gegevens toegankelijk maakt voor een groot publiek in een openbare cloud, moet je bewust zijn van de kosten voor uitgaand data verkeer! De hoeveelheid data (=gebruik * grootte) kan hoge kosten met zich meebrengen en je moet daarbij rekening houden met de business case.

Als je meer in detail op dit onderwerp wilt ingaan, raadpleeg dan GitHub waarin een uitgebreid artikel gepubliceerd is over het plaatsen van gegevens¹¹.

¹¹ https://pages.github.ibm.com/IBMAoT/i-hc_data_placement

10. SORTS OF DATA



Sorts of Data

Je kunt data op allerlei manieren classificeren en categoriseren. Ook binnen het Sol domein zijn er verschillende toepassingsgebieden. Een data scientist heeft vooral behoefte aan ruwe, onbewerkte data. Een BI gebruiker wil juist gecureerde en genormaliseerde data hebben om rapporten te kunnen bouwen. In dit hoofdstuk geven we een aanzet om na te denken over die toepassingsgebieden. Elk gebied heeft zijn eigen karakteristieken. Denk maar eens aan:

- Van gestructureerd naar ongestructureerd,
- van operationeel tot analytisch,
- van stamgegevens tot referentie-informatie,

- van real-time tot historische informatie,
- van ruwe data tot genormaliseerde data.

De volgende stap is het bepalen van de beste technologie die dat specifieke gegevensformaat ondersteunt. Een referentietabel kan bijvoorbeeld bestaan uit ‘name value pairs’ die het beste kunnen worden opgeslagen in een key-value database zoals Redis of etcd, terwijl je master data kan worden opgeslagen in een relationele database zoals DB2 of PostgreSQL.

Zorg ervoor dat je architectuur governance proces op orde is zodat je kunt voorkomen dat er een te grote diversiteit aan oplossingen ontstaat voor gegevensopslag.

Opdracht:

Maak een overzicht van Sorts of Data jij onderkent en bepaal welke technologie hier het beste bij past.

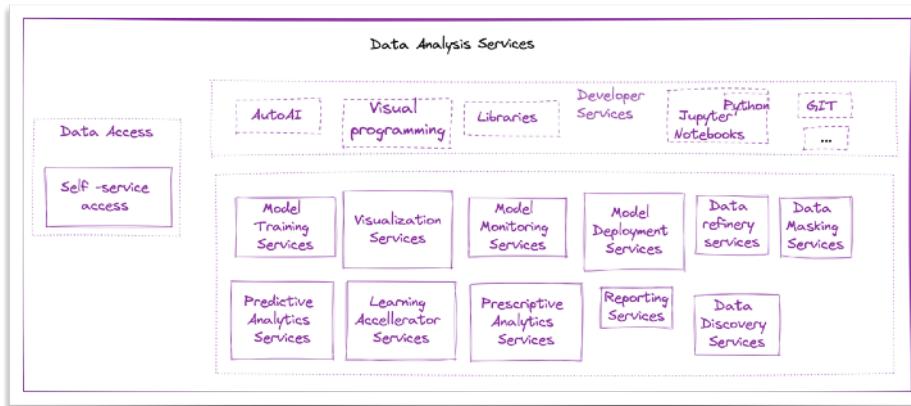
Master Data	Master Data zijn gegevens gerelateerd aan de basisregistratie. Kijk naar de registratie van burgergegevens zoals naam en adres of de registratie van voertuigen. Bij voorkeur is er slechts één bron van de Master Data. De zogenaamde ‘single source of truth’. Om allerlei kopieën te voorkomen, wordt Master Data via API's beschikbaar gesteld. Master Data behoort naar onze mening bij een Systems of Insight, ook al zullen ze veelvuldig worden gebruikt door Systems of Record.
-------------	--

Reference Data	Reference Data is vergelijkbaar met Master Data, maar het zijn meer statische gegevens. Land- en taalcodes zijn voorbeelden. Maar ook de prijzen voor energie worden als referentie data beschouwd ook al kunnen deze per uur verschillen.
Enterprise Content Data	Enterprise Content Data is ongestructureerde of semi-gestructureerde data. Grote hoeveelheden gegevens zoals documenten, e-mails, video's, spraakopnamen en afbeeldingen worden allemaal als "content" beschouwd. Om ervoor te zorgen dat systemen deze gegevens begrijpen, worden metadata gebruikt die bij de inhoud horen. Een enterprise content management database (ECM) bestaat uit twee delen: de metadata, gebruikt als index, en de inhoud zelf. AI wordt veelvuldig gebruikt om de inhoud te interpreteren.
Operational Data	Operationele Data bevat de operationele data van de Systems of Insight. Dit kan een relationele database zijn, een noSQL-database of zelfs de etcd-database van de Kubernetes-omgeving waar alle configuratiegegevens zijn opgeslagen.

Historical Data	<p>We moeten ervoor zorgen dat bij de operationele systemen en enterprise content systemen historische informatie verwijderd wordt zodat de systemen blijven presteren. Hoewel deze gegevens misschien waardeloos zijn voor de operationele systemen, kan de waarde van historische gegevens voor de data scientist enorm groot zijn. Daarom worden gegevens bewaard voor analytische doeleinden van welke aard dan ook. Vaak wordt gevoelige informatie zoals burgerservicenummers geanonimiseerd of gepseudonimiseerd bij deze analytische toepassingen.</p>
Raw Data	<p>Raw Data bevat kopieën van operationele gegevens die door de Systems of Insight kunnen worden gebruikt. De kopie wordt gemaakt om operationele impact op de Systems of Record te voorkomen. Het meest voorkomende formaat zijn relationele databases, maar het kunnen ook andere soorten gestructureerde gegevens zijn, zoals NoSQL-databases. Deze ruwe data kan worden gebruikt om data-warehouses en datamarts te voeden. Deze ruwe datakopieën worden doorgaans opgeslagen in een zogenaamde staging area (tussenliggende/tijdelijke resultaten). Ruwe data wordt ook gebruikt om analyses op los te laten.</p>
Sandbox	<p>De sandbox is een veilige speeltuin voor de data scientist. Hij of zij kan geen kwaad doen en ze kunnen naar hartenlust met de data experimenteren.</p>

Conformed Data	De conformed data bevat samengestelde gegevens die klaar zijn voor gebruik. Dit kan een data-warehouse, een datamart of de trainings dataset zijn voor een model.
----------------	---

11. DATA ANALYSIS SERVICES



Data Analysis Services

Data Analysis Services kan worden beschouwd als de kern van de diensten die een data-gedreven organisatie nodig heeft. Het kan diepgaande inzichten geven door analyse van de onderliggende data. Van het reflecteren op gebeurtenissen uit het verleden vanuit historische data tot voorspellende inzichten over wat er gaat gebeuren.

AutoAI	AutoAI is een mogelijkheid voor leken of data scientists om snel te kunnen starten met het maken van AI modellen. Het kan gebruikt worden door data scientists die geen programmeer achtergrond hebben.
--------	---

Visual Programming	Visual Programming is een tool die data scientists ondersteunt bij visualiseren van een datastream. Door bronnen en bewerkingen naar een canvas te slepen, kan je op een eenvoudige manier modelleren. Tools zoals SPSS Modeler zijn voor de meeste data scientists bekende tools voor visuele datavoorbereiding en modellering.
Jupyter Notebooks	Het typische hulpmiddel voor de data scientist om modellen te maken is het gebruik van een document formaat (notebook). Het kan gebruikt worden met meerdere ontwikkeltalen, maar de naam Jupyter komt van drie populaire talen: Julie, Python en R.
Libraries	Analytische en statistische (open source) bibliotheken die door data scientists voor hun programma's gebruikt kunnen worden.
Python	De ontwikkeltaal voor de programmeur data scientist. Een andere veelgebruikte taal is R. Deze is ontwikkeld voor statistische en data-analytische doeleinden. We vermelden de meest populaire taal, maar deze dienst had ook 'analytische programmeertaal diensten' kunnen heten. Je kunt dit gerust vervangen door de talen die in jouw organisatie worden gebruikt.

GIT	<p>GIT is een opslagplaats van broncode van welke aard dan ook. Verschillende versies kunnen in een GIT-repository worden bewaard. In een cloud native omgeving is het gewoonlijk om applicaties vanaf <i>scratch</i> te bouwen. Zo kom je niet in situaties terecht waarbij de broncode van een draaiende module niet gevonden kan worden. Voor betrouwbare AI is het belangrijk om zowel de bronapplicaties als de brondaten bij te houden. GIT ondersteunt deze vorm van documentatie.</p>
Data Masking Services	<p>Gegevensmaskering voor gestructureerde gegevens betekent anonimisering of pseudonimisering van de gegevens wanneer deze worden weergegeven aan doelgroepen die niet de rechten hebben om deze te lezen. Vaak wordt dit gebruikt om te voldoen aan de AVG. Er wordt dan een sterretje weergegeven in plaats van een bankrekening, of er wordt een willekeurige bankrekening gegenereerd die niet bestaat.</p>
Model Training Services	<p>De service om het model te trainen door er gegevens in te voeren.</p>
Visualisation Services	<p>Deze services visualiseren de uitkomst van de modellen op een gebruiksvriendelijke manier. Verschillende visualisaties kunnen verschillende boodschappen overbrengen. Het kiezen van de juiste visualisatie zorgt ervoor dat de uitkomst van een model snel kan worden geïnterpreteerd door een gebruiker. Een visualisatie zou bijvoorbeeld een wereldkaart kunnen zijn maar ook een piechart.</p>

Model Monitoring Services	Bewaakt het geïmplementeerde model door te valideren of het model nog steeds binnen gedefinieerde grenzen functioneert. Het meet vertrouwen, drift, eerlijkheid, bias, enz.
Model Deployment Services	Ondersteuning van de modeloverdracht naar productie omgeving. Dit wordt ervaren als een complexe en tijdrovende activiteit, maar met de model deployment service kan de implementatie van modellen naar productie worden geautomatiseerd.
Data Refinery Services	Data scientists hebben toegang tot grote hoeveelheden data. Data refinery service ondersteunt de data scientist bij het maken van selecties, transformaties en manipulatie van gegevens die nodig zijn voor een specifieke analyse.
Data Discovery Services	Data Discovery Services zijn intelligente componenten die typen gegevens ontdekken door de inhoud ervan te lezen. Het kan b.v. een e-mailadres vinden en classificeert dat als persoonlijke data. Data Discovery Services kunnen honderden verschillende datatypen identificeren en de data-operator ondersteunen bij het bouwen van een datacatalogus. Dit wordt gebruikt door de data engineer en de data curator om de data beschikbaar te maken voor gebruikers.
Predictive Analytics Services	Het biedt voorspellende inzichten zoals trends en afwijkingen op basis van de modellen. AIOps gebruikt deze diensten ook om mogelijke incidenten te voorspellen door normaal gedrag te observeren en afwijkingen te detecteren.

Prescriptive Analytic Services	Het geeft aan welke actie moet worden ondernomen op basis van de inzichten die modellen genereren. AIOps gebruikt dit ook om acties te ondernemen om incidenten te voorkomen. Op deze manier kunnen systemen zichzelf ‘genezen’, zogenaamde zelfherstellende systemen. Omdat de data gedreven organisatie in de loop van de tijd waarschijnlijk duizenden microservices zal implementeren, worden deze services steeds belangrijker.
Learning Accelerator Services	Machine Learning Accelerator is een AI-infrastructuur dienst die het GPU gebruik optimaliseert.
Reporting Services	De reporting service biedt de mogelijkheid om op verzoek dag-, week-, kwartaal- en maandrapportages of ad-hoc rapportages en dashboards te creëren (Business Intelligence).
Self Service Access	Een Self Service dienst is een belangrijk onderdeel van een Data Fabric (zie hoofdstuk Data Fabric). De data engineer maakt de data beschikbaar via een self service portaal. De data scientist gebruikt dit portaal om een selectie te maken van de noodzakelijke data en hoeft niet iedere keer terug te vallen op een proces om data te verkrijgen.

11.1. AI GOVERNANCE STAAT HAAKS OP HET AGILE MANIFEST

Begrijpelijk dat je je wenkbrauwen fronst. Wat bedoelen ze met deze uitdagende titel? Laten we het proberen uit te leggen. Het agile manifest luidt: **“Werkproduct boven uitgebreide documentatie”**, er staat ook: **“Individuen en interactie boven processen en hulpmiddelen”**.

Beide statements zijn niet echt een goed idee als we het hebben over de betrouwbaarheid van AI en de AI-regelgeving van de EU. En aangezien er maar vier uitspraken in het agile manifest staan, hebben we het dus over 50% van het manifest dat de plank flink misslaat. Dat doet pijn!

Vanuit het oogpunt van betrouwbaarheid zou het dus zinvol zijn om het agile manifest zorgvuldig te herzien, omdat het je 30 miljoen euro of 6% van je wereldwijde bedrijfsinkomsten kan kosten als je niet voldoet aan de EU-AI-regelgeving. Hebben we je aandacht?

Het doel van de regelgeving is om ervoor te zorgen dat AI eerlijk gebruikt wordt. Deze regelgeving is van toepassing op elk AI-systeem dat in de EU wordt gebruikt. Ze zijn alleen van toepassing op situaties waarin AI wordt blootgesteld aan het publiek, bijvoorbeeld een chatbot die je adviseert jouw hete, op 240 V aangesloten haardroger af te koelen onder de koudwaterkraan... Sommige toepassingen van AI zijn binnen de EU helemaal niet toegestaan, bijvoorbeeld social scoring. Bovendien liggen applicaties die gezichtsherkenning toepassen erg gevoelig, daarom zijn bedrijven als Microsoft en IBM gestopt met deze diensten.

Deze regelgeving is voor gebruik binnen je eigen organisatie niet van toepassing. Dus als je je eigen organisatie helemaal kapot maakt, kan het de EU kennelijk niks schelen. Het is aan jou om te voorkomen dat AI verkeerd wordt ingezet binnen jouw organisatie. Daarom kan het een goed idee zijn om de regelgeving als leidraad te gebruiken om je eigen organisatie tegen ernstige misstappen te beschermen.

Bij het uitbrengen van de eerste versies van dit boek was de regelgeving nog niet van kracht maar hebben we het wel geadresseerd. Gelukkig hebben die lezers ons advies ter harte

genomen en deze regelgeving geïmplementeerd, maar jij, die het nu past leest, je bent te laat!

Dus verlies geen moment en begin met een grondige aanpak door de governance- en compliance diensten te implementeren die we hebben voorgesteld om governance in je eigen organisatie in te richten. Dit is ook de reden waarom we enkele zinnen aan ModelOps hebben besteed en waarom we de model monitoring service hebben beschreven om b.v. model drift te detecteren. Al governance is onderdeel van het architectuur domein Data Analysis Services. Data Governance is onderdeel van Data Governance en Compliance Services.

We hebben vijf ‘vertrouwenspijlers’ gedefinieerd waaraan jouw AI-oplossing volgens ons moet voldoen:

1. Eerlijkheid
2. Uitlegbaarheid
3. Robuustheid
4. Zekerheid
5. Governance

Houd er rekening mee dat er veel discussie is over dit onderwerp en dat er ook veel goede principes zijn vastgelegd die je kunt overnemen.

Hier zijn als voorbeeld de drie AI principes van IBM:

1. Het doel van AI is om de menselijke intelligentie te vergroten

Het doel van de door IBM ontwikkelde en toegepaste AI- en cognitieve systemen is het vergroten – en niet vervangen – van de menselijke intelligentie. Onze technologie is en zal worden ontworpen om de menselijke capaciteiten en potentie te vergroten en uit te breiden. Bij IBM zijn we van mening dat AI ons allemaal beter moet maken in ons werk en dat iedereen de voordelen van het AI-tijdperk moet voelen en niet alleen de elite. Daartoe investeren we in initiatieven om de mondiale

beroepsbevolking te helpen de vaardigheden te verwerven die nodig zijn om samen te werken met deze technologieën.

2. Data en inzichten zijn van de eigenaar

De data is van onze klanten, en hun inzichten die zijn gecreëerd zijn van onze klanten. Klantgegevens en de inzichten die in de cloud van IBM of vanuit de AI tools van IBM worden geproduceerd, zijn eigendom van de klanten van IBM.

3. Nieuwe technologie, inclusief AI-systemen, moet transparant en verklaarbaar zijn

Wil het publiek AI kunnen vertrouwen, dan moet het transparant zijn. Technologiebedrijven moeten duidelijk zijn over wie hun AI-systemen traint, welke gegevens bij die training zijn gebruikt en belangrijker nog, wat er in de aanbevelingen van hun algoritmen is verwerkt. Als we AI willen gebruiken om belangrijke beslissingen te nemen moet dit verklaarbaar zijn.

Dit geldt voor bedrijven zoals IBM maar ook voor overheidsorganisaties. Wij zijn van mening dat het databeleid van de overheid eerlijk en rechtvaardig moet zijn en prioriteit moet geven aan openheid.

Dit zijn geen uitspraken die je alleen maar opschrijft. Dit zijn uitspraken die je binnen je organisatie moet implementeren en moet naleven! De overheid heeft het algoritmeregister in het leven geroepen om algoritmes verantwoord en transparant toe te passen.

11.2. HET ALGORITMEREGISTER

De regering wil dat de overheid algoritmes verantwoord gebruikt. Mensen moeten erop kunnen vertrouwen dat algoritmes voldoen aan de waarden en normen van de samenleving. En er moet uitleg zijn over hoe algoritmes werken. Wanneer de overheid

open is over algoritmes en hun toepassing, kunnen burgers, organisaties en media haar kritisch volgen en controleren of ze de wet en de regels volgt.

Algoritmes.overheid.nl is het centrale Algoritmeregister van de Nederlandse overheid. Hier publiceren overheden informatie over hun algoritmes op één plek, zodat het doorzoekbaar is voor iedereen. Het Algoritmeregister is een van de acties om algoritmes te reguleren uit de Werkagenda Waardengedreven Digitaliseren.

11.3. VERANTWOORD ALGORITMEGEBRUIK

De volgende trajecten dragen bij aan verantwoord algoritmegebruik:

1. Het Algoritmeregister helpt om algoritmes vindbaar te maken, beter uit te leggen en hun toepassing en impact begrijpelijk te maken.
2. De Algoritmetoezichthouder (de Autoriteit Persoonsgegevens) coördineert de controle op algoritmes: voldoen de algoritmes van de overheid aan alle regels die daarvoor gelden?
3. Het ministerie van BZK werkt aan het Algoritmekader. Dat maakt voor overheden duidelijk welke eisen er gelden voor algoritmes en hoe ze ervoor kunnen zorgen dat hun algoritmes daaraan kunnen voldoen.
4. Wetgeving: er komt een wettelijke verplichting voor transparantie van algoritmes. Dat is aangekondigd in de kamerbrief van december 2022¹².

11.4. DOELEN VAN HET ALGORITMEREGLER

Wat willen we bereiken met het Algoritmeregister? De doelen van het Algoritmeregister hangen met elkaar samen. Hieronder

¹² <https://www.rijksoverheid.nl/documenten/kamerstukken/2022/12/21/kamerbrief-over-het-algoritmeregister>

worden de doelen kort toegelicht. Een uitgebreidere toelichting op de doelen vind je in de Handreiking Algoritmeregister¹³.

1. Vertrouwen in de overheid vergroten

De overheid kan alleen effectief maatschappelijke meerwaarde bieden, als er vertrouwen in de overheid is en mensen zich gehoord en betrokken voelen. Het Algoritmeregister moet helpen het vertrouwen in de overheid te verbeteren.

2. Versterken van de positie van burgers en bedrijven

Met het Algoritmeregister krijgen burgers en bedrijven meer informatie in handen over hoe processen van de overheid werken en welke keuzes zijn gemaakt.

3. Meer duidelijkheid over algoritme- en AI-gebruik (Demystificatie)

Voor veel mensen is onduidelijk wat een algoritme is en hoe deze effect heeft op zijn of haar leven. Het Algoritmeregister maakt inzichtelijk waar en hoe de overheid algoritmes en AI gebruikt.

4. Verantwoord handelen

De overheid moet op een verantwoorde manier handelen. Het Algoritmeregister maakt het handelen van de overheid meer openbaar.

5. Controleerbaarheid van de overheid vergroten

Als de overheid laat zien wat ze doet, kunnen burgers en organisaties haar beter controleren. Het Algoritmeregister ondersteunt deze controle door de maatschappij.

6. Vergroten uitlegbaarheid

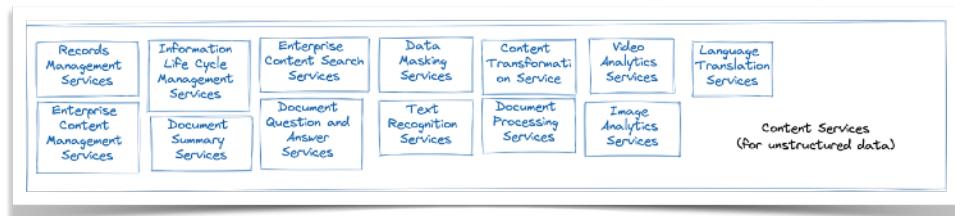
Het vullen van het Algoritmeregister dwingt de overheid om haar besluitvormingsproces en de toepassing van algoritmes begrijpelijk uit te leggen. Daarnaast kunnen burgers hierdoor aangeven of een uitleg duidelijk is en of zij de werking hetzelfde ervaren.

¹³ <https://algoritmes.pleio.nl/wiki/view/19bb6e9e-7a97-43d5-bef3-b1d66e59f4ff/handreiking-algoritmeregister>

7. Transparanter zijn over het handelen

Om openheid en controleerbaarheid te kunnen bieden, moet de overheid transparant zijn. Het zorgvuldig vullen van het Algoritmeregister is een van de manieren om dat te doen.

12. CONTENT ANALYSIS SERVICES



Content Analysis Services

Wat Data Analysis Services zijn voor gestructureerde data, zijn Content Analysis Services voor ongestructureerde data. Tijdens onze gesprekken met experts hebben we goed inzicht gekregen in het verschil tussen het gebruik van content en gestructureerde data. Een paar kenmerken:

Structured	Unstructured
Goed voor voorspellingen	Gebruikt als bewijsmateriaal
Voorschrijvend voor het bepalen van de volgende stappen	Knowledge base
Reports en Business Intelligence	Content + Structuur (Metadata)
Grote volumes, kleine hoeveelheden	Grote files

Text Recognition Services	Tekstherkenningsdiensten herkennen (handgeschreven) tekst. Vaak gebruikt door de Input Management Services om papieren formulieren om te zetten in een digitaal formaat met de bijbehorende metadata.
Records Management Services	Records Management is de implementatie van Information Life Cycle Management voor ongestructureerde data. Content kan onder governance of archiefbeheer vallen. Op het moment dat aan de content een record wordt toegewezen, wordt het aanmaken, gebruiken, bewaren en verwijderen van die content toegewezen en afgehandeld door het archiefbeheer.
Enterprise Content Search Services	Enterprise Content Search Services doorzoekt een federatie van content repository's, zowel in metadata als in de content zelf. Voor de content zelf kan het gebruik maken van diensten zoals beeld- en video analyse diensten.
Enterprise Content Management Services	ECM is voor ongestructureerde data wat een RDMS is voor gestructureerde data. Het is nauw verwant aan Enterprise Content Data (zie Sorts of Data). Het verwerkt alle soorten activiteiten die op de gegevens worden uitgevoerd, zoals bekijken, eenvoudig zoeken, metadata koppelen aan inhoud, enz.

Document Summary Services	<p>The Document Summary Service is in staat om één of meer documenten voor jou samen te vatten. Naast die samenvatting kan het ook aangeven wat het sentiment is. Het is moeilijk om goede samenvattingen te maken. Met de komst van GenAI lijkt dit steeds beter te lukken. Veel van deze diensten zijn alleen beschikbaar in de cloud en dat stelt weer strenge eisen aan privacy (AVG). Als een document al moeilijk is, hoe ga je een video samenvatten? Het zou ook mooi zijn als mijn vakantiefoto's samengevat worden! De grote vraag bij deze dienst om samen te vatten blijft altijd: "Stond er niet iets in het document dat over het hoofd is gezien dat toch belangrijk was?".</p>
Document Question and Answer Services	<p>Document question and answer services leest en interpreteert het document zodat je er vragen aan kunt stellen, die vervolgens door de dienst worden beantwoord. Een voorbeeld is GAAMA. Verveeld door het lezen van al deze dingen? Ga naar internet, zoek naar "GAAMA is a Natural Language Understanding technology", ga naar de zijkant, selecteer het tabblad "aangepast" en laad de PDF van dit boek (kopiëren - plakken) en stel de vraag: "wie zijn de auteurs? " Dat zijn wij! Aangenaam kennismaken! Probeer dan eens een minder serieuze vraag als 'wat is een data-gedreven organisatie'. Wauw, die kent het! Het weet zelfs wie Data Mesh heeft uitgevonden! Wat deze service nuttig maakt, is dat deze de zekerheidsfactor laat zien en de regel waar het antwoord is gevonden om te valideren of het antwoord juist is.</p>

Data Masking Services	Data Masking Services voor ongestructureerde data maskeert gegevens op basis van het profiel van de gebruiker. Voorbeelden hiervan zijn het maskeren van kentekenplaten. Vervang vloeken door piepjess, vervang namen (AVG) door rollen. Zwarte blokken (zwartlakken) zijn ook een vorm van gegevensmaskering in documenten.
Content Transformation Services	Content Transformation Services kan de transformatie van het ene dataformaat naar het andere uitvoeren. Het transformeert bijvoorbeeld een spraakbestand naar tekst, een videobestand naar afbeeldingen of een PDF naar gegevens. OCR kan gezien worden als een specifieke vorm van Content Transformatie.
Document Processing Services	De Document Processing Services neemt een document, interpreteert de context van dat document en zorgt dat deze context in het juiste bedrijfsproces wordt verwerkt. Je hebt bijvoorbeeld een factuur ontvangen, de Document Processing Services leest de factuur en herkent het te betalen bedrag, de afzender, het bedrag aan belasting etc. en stuurt het juiste bedrijfsproces aan met de bijbehorende informatie om de betaling te verwerken. Document processing service kan worden gebruikt door invoerbeheer voor documenten. Als delen van het formulier handgeschreven of "tekst in een afbeelding" zijn, kunnen tekstherkenning services worden gebruikt om die invoer te vertalen naar machinaal leesbare tekens.

Video Analytics Services	Deze dienst doet voor video hetzelfde als wat de beeldanalyse services voor afbeeldingen doen. Je levert bijvoorbeeld een dashcam video aan en deze haalt automatisch het videofragment op waar het ongeval plaatsvindt. Of het verzamelt de gezichten van mensen die geweld gebruiken in een stadion.
Image Analytics Services	Met behulp van AI analyseert deze service afbeeldingen. Beeld analyse omvat het verwerken van een beeld in fundamentele componenten om betekenisvolle informatie te extraheren. Beeld analyse kan taken omvatten zoals het vinden van vormen, het detecteren van randen, het verwijderen van ruis, het tellen van objecten en het berekenen van statistieken voor textuur analyse of beeldkwaliteit. Afhankelijk van je context wil je dit misschien opsplitsen in meerdere services om de nauwkeurigheid van de analyse te vergroten. Een voorbeeld zou de "Franse kenteken herkenningsdienst" kunnen zijn. Het zou gaan om een beelddienst die speciaal is getraind voor het herkennen van Franse kentekenplaten. Vaak vind je in een data-gedreven organisatie veel van dit soort analytische diensten.
Language Translation Services	De language translation services vertalen inhoud van de ene taal naar de andere. Dat kunnen tekst zijn, maar ook gesproken opnames. Hoe zit het met het vertalen van de tekst in afbeeldingen? Hoe zou jij die vertalen?

Auto Classification Services	De Auto Classification Services kunnen de inhoud van je data lezen en de metadata voor die inhoud bepalen en ook een beveiligingsclassificatie aan je document toekennen. Bijvoorbeeld wanneer er persoonlijke gegevens zoals een e-mailadres in het document worden aangetroffen.
------------------------------	--

12.1. EEN CONTENT SERVICES PLATFORM

Misschien zegt het je wat, maar misschien ook niet. Een Content Services Platform (CSP) is een relatief nieuw begrip dat Gartner geïntroduceerd heeft voor een omgeving waar met “content” gewerkt wordt.

In den beginne ... was er een “Document Management System (DMS)” dat werd “Enterprise Content Management (ECM)” en nu is dit geëvolueerd naar een Content Services Platform (CSP).

Gartner “definieert” een CSP als volgt:

Content services platforms (CSPs) are foundational for ***managing and utilizing*** content within an organization. CSP technology enables employees to retrieve and work with ***content in a modern and seamless way across devices and organizational boundaries***. Core CSP functionalities include ***content capture, creation, consolidation, processing, and retention*** to support personal, team, departmental and enterprise business operations.

Verder heeft Gartner in een rapport¹⁴ 15 kritische capabilities genoemd waar een CSP aan moet voldoen. Laten we deze eens langslopen en vergelijken met de services die wij geïdentificeerd hebben.

Collaboration. Dit is al gelijk een interessante capability. Aan de ene kant wil je je content onder controle houden (Information Governance), aan de andere kant wil je juist zoveel mogelijk ad-hoc samenwerken. Wanneer stopt het één en begint het ander? Wat ons betreft bevindt collaboration (dat is veel meer dan het delen van een document) zich vooral in het Workspace domein. Het gaat dan over samen aan één object werken, versturen per email, delen via social media, etc. Dat hebben wij niet geadresseerd als onderdeel van de Sol.

Probeer dat maar eens onder governance te doen! Je kunt er wel voor zorgen dat de content processing onderdeel is van workflow en processen. Op dat moment lever je echter wel veel vrijheid in t.a.v. collaboratie. In het volgende hoofdstuk gaan we hier iets uitgebreider op in omdat het een belangrijk theme is binnen de Informatie Huishouding (IHH).

Productivity. Om een goede productiviteit te bereiken moet het platform goed geïntegreerd zijn en moet je op het platform makkelijk (no-code/low code) applicaties kunnen bouwen en snel deployen. De CI/CD pipeline services die wij genoemd hebben kunnen hierbij helpen.

Capture. Deze capability hebben wij opgenomen onder de naam “Input Management Services” en gepositioneerd in het “Business Process Support services” architectuur domein. Wanneer je een aantal content analytics services (text, image, video) hieraan toevoegt heb je een prachtige set van capture capabilities.

¹⁴ [1] <https://www.m-files.com/what-are-gartner-s-15-critical-capabilities-of-content-services-platforms-csp-3/>

Search. Deze capability hebben wij opgenomen als “Enterprise Content Search Services”.

Platform. Het idee van een platform is dat je een fundament hebt waar je op voort kunt bouwen. Daarom hebben wij in dit boek een container platform als basis (Application Runtime). Je kunt daarop standaard applicaties deployen maar ook zelf applicaties (microservices) toevoegen.

Administration. Uiteraard wil je allerlei administratie kunnen doen. Bijvoorbeeld security. Dit vinden wij logischerwijs bij elk van de capabilities horen en daarvoor hebben wij dus geen extra service opgenomen.

Analytics en rapportages. Met analyses en rapportages kun je inzicht krijgen in datgene wat er gebeurt op het CSP. Omdat een CSP ook workflows en processen kent, is de analyse niet beperkt tot de content, maar ook wordt er gekeken naar wat er met de services gebeurt. Dit valt onder de services Business Process Monitoring.

Metadata en classificatie. Hoewel dit in een zin wordt genoemd is er een nuance aan te brengen tussen metadata en classificatie. De classificatie van een document (bijv. CIA classificatie) wordt veelal gezien als metadata. Het platform moet zowel de mogelijkheid bieden om deze informatie op te slaan, maar ook mogelijkheden om deze automatisch te genereren, bv. door de inzet van AI technologie.

Workflow en processen. In het architectuur domein Business Process Support services hebben we zowel case management services als business proces services opgenomen. Ook hier voegt Gartner deze twee samen. Wij zien een belangrijk verschil tussen beide.

Business Application Enablement. Het moet mogelijk zijn om business applicaties te integreren in het CSP. Dit kan bijvoorbeeld middels een API zoals we dat hebben benoemd in het Integration Services architectuur domein. Een business applicatie kan ook een workflow of een proces zijn. In onze definitie van case management en Business Process Services is er ruimte voor business applicaties.

Information Governance. Information Governance wordt voor een groot gedeelte afgedekt door de services Records Management en Data Lifecycle Management. Het architectuur domein Data governance en compliance biedt ook een aantal services hieromtrent, maar is wat meer gericht op de gestructureerde data.

Content management. Deze capability noemde wij in het verleden dus Enterprise Content Management. Dat is dus onze Content Management Service die je terugvindt in het domein Content Analysis Services.

Delivery Model. Zie voor het delivery model de CI/CD Pipeline Services. Ook Gartner geeft dit onderwerp dus extra aandacht. Neem niet zomaar aan dat het delivery model voor een CSP gelijk is aan de delivery modellen die jij al gebruikt voor je SoR of SoE.

Content Security. Content, data of information security is ontzettend belangrijk. We zien in de praktijk dat veel bedrijven dit onvoldoende ingericht hebben. De bedrijven die dit wel hebben ingericht kiezen voor een enterprise oplossing. De Content van het CSP wordt dus beveiligd door de Enterprise oplossing. Het CSP zelf biedt die oplossing niet, maar moet ermee kunnen integreren. In ons hoofdstuk over Data Security Services hebben we heel wat geschreven over dit onderwerp.

System Security. Misschien zou dit beter platform security kunnen heten. Het gaat hier om applicatie toegang. Wij hebben dit niet expliciet beschreven omdat dit onderdeel is van de bestaande omgeving, Je CSP moet integreren met je access management oplossing.

Het is een behoorlijk lijst van capabilities die horen bij een CSP. Dus heel wat uitgebreider dan een Document Management System! Op de één of ander manier komen al deze capabilities terug in onze blauwdruk. Denk eens na over het volgende:

Wanneer een CSP voor jou relevant is adviseren we je om de volgende opdracht uit te voeren:

Pak de A0 poster en maak een mapping van alle capabilities die jij (voor jouw organisatie) relevant vindt. Het resultaat is een mooi begin voor een CSP referentiearchitectuur!
Had je al services geïdentificeerd? Zijn er die wij gemist hebben?

12.2. COLLABORATIE

Wanneer wordt informeel formeel?

Ideeën ontstaan op vele manieren. Het kan een opwelling zijn, een ongenoegen dat je wilt oplossen of het resultaat van een gestructureerde ideation workshop. Ideeën zijn een creatief resultaat dat je niet kunt en ook meestal ook niet wilt vangen in een tool.

Veel ideeën zullen in de praktijk stranden in schoonheid. Er zijn meestal meer ideeën dan de hoeveelheid geld, tijd en resources die beschikbaar zijn om al deze ideeën uit te werken. Maar er zullen ook ideeën overleven en deze zullen we verder gaan uitwerken. We maken een plan, schrijven een concept en

bespreken dit met anderen. Op een zeker moment is het formeel en het idee zou een principe kunnen zijn, een beleidsstuk of zelfs een wet. Ergens in dit traject moet het creatieve, informele omgezet worden naar structuur waarbinnen onderwerpen zoals governance, archivering en records management een belangrijke rol gaan spelen. Als je te snel start met het formaliseren van het proces dan blokkeer je creativiteit en snelheid, ben je te laat dan ben je context en controle kwijt waardoor de beslissing en het proces om tot die beslissing te komen continue ter discussie worden gesteld. Het kan je uiteindelijk veel tijd kosten om het

Een voorbeeld van een informeel proces. De gemeente heeft besloten om een voorstel aan te nemen om een zonnepark te plaatsen op de Bloemenweide. Het idee was ontstaan doordat Kees, een gemeenteraadslid, op een zondagochtend langs de Boterbloemweg fietste en dacht dat dit wel een mooi lokatie zou zijn voor een zonnepark. Hij stuurde een email naar Frits, zijn collega gemeenteraadslid, die dit wel een mooi idee vond. Ze gaan samen zitten en laten onderzoeken of er voldoende bekabeling is en wat het bestemmingsplan is. Kosten: twee dagen tijd voor twee personen en 100 euro voor het opvragen van de gegevens bij het kadaster en de energieleverancier. Ze schrijven een voorstel en bespreken dit met een tweetal wethouders. Het wordt aangepast en ze dienen het voorstel in voor de gemeenteraadsvergadering...

Wanneer is het formele proces gestart? Je zou kunnen zeggen op het moment dat het voorstel formeel wordt ingediend. Maar ze hebben wel geld uitgegeven en tijd besteed aan het maken van hun plan, moet daar ook verantwoording voor worden afgelegd? Zijn de aanpassingen die gedaan zijn door input van de wethouders relevant voor het archief?

(weer) uit te leggen en wanneer het niet helemaal duidelijk is kan het makkelijk tot reputatieschade leiden.

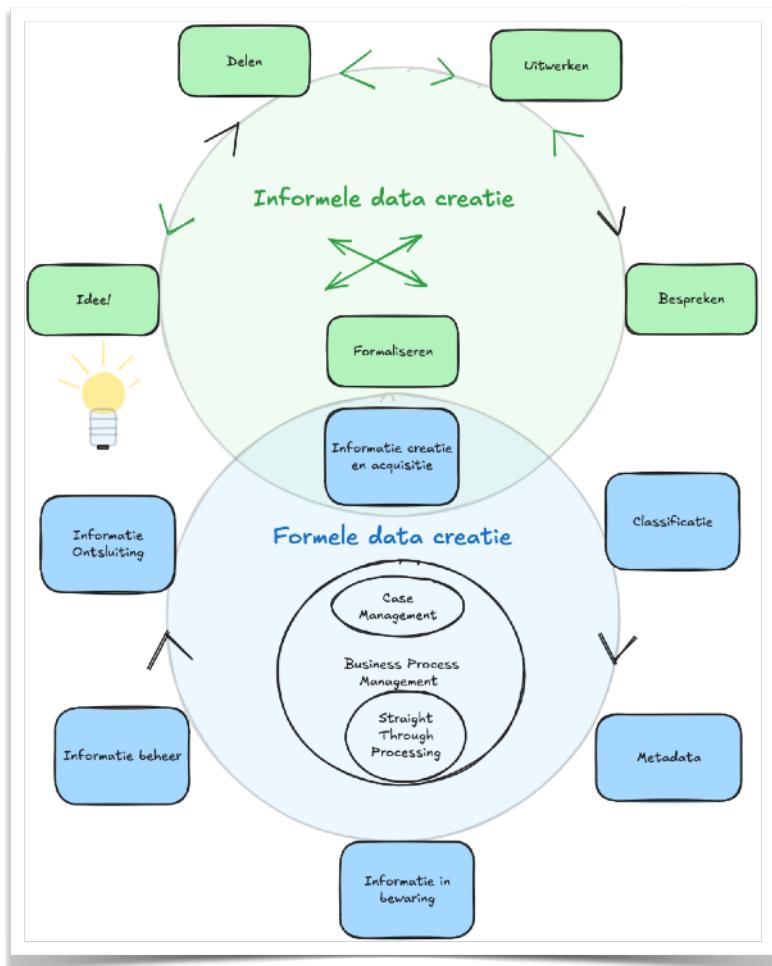
De karakteristieken van een formeel en informeel proces geven weer waarom de overgang van informeel naar formeel meer is dan allen maar een knopje omzetten:

Informeel	Formeel
Snelheid om tot resultaat te komen	Alles wordt vastgelegd
Veel (korte) interacties	Formele afstemming en vastlegging waardoor je op ieder moment in de tijd de status kan achterhalen
Out of the box (de wet is even niet van toepassing)	Binnen de box blijven
Gesprek bij de koffiemachine	Gesprek in een vergadering
Een apje, een telefoontje	Notulen, formeel proces
Spontaan uitflappen	Weloverwogen politiek correcte communicatie
Emotie (een grap, lachen, verdriet)	Ratio, functioneel emotioneel
Mening en opinie	Feiten
Cultuur	Structuur
Achterkant van het bierfiltje	IT systeem met backup en resilience
De kroeg	Het kantoorpand
De pionier (constante aanpassing)	De administrateur (structuur, orde in chaos)

Het is niet alleen lastig om het moment te bepalen dat je overgaat van informeel naar formeel, maar het is ook een

cultuurverandering. Je moet in één keer ander gedrag gaan vertonen.

De aanzet voor IHH was het niet tijdig en accuraat documenteren en onder controle hebben van besluitvorming. Overweeg dus heel goed wanneer je de documentatie mbt besluitvorming gaat behandelen als officieel. Wanneer wordt het relevant om in de toekomst hierover informatie op te vragen?



Rich Robinson schets in zijn boek “All Change”¹⁵ 4 fasen: Dream, Discover, Design en Deploy. In de droom fase worden ideeën bedacht, in de discover fase worden ze uitgewerkt (prototype) en in de design fase wordt een idee uitgeprobeerd (pilot) en vervolgens wordt het in productie genomen. Je zou kunnen zeggen dat de eerste twee fasen bij het informele proces horen en de laatste twee bij het formele. Maar in de praktijk is het een iteratief proces en zul je meestal de processen niet kunnen scheiden.

12.3. RECORDS MANAGEMENT

We produceren met zijn allen iedere dag weer enorme hoeveelheden data. Het meeste van deze data kan in verloop van tijd weer de prullenbak in maar sommige data is zo belangrijk dat

De US Federal Records Act definieert een “record” als volgt:

“...recorded information, regardless of medium or characteristics, made or received by an organisation that is evidence of its operations and has value requiring its retention for a specific period of time.”

we precies willen weten wat en wanneer er iets met die data gebeurt. We besteden dat beheer uit aan een zogenaamde Records Manager. De data, bijvoorbeeld een belangrijk contract wordt bestempeld als een “record”. De records manager zal vanaf dit moment dit contract bewaken alsof zijn leven ervan afhangt. Zelfs als het contract ten einde is en hij het contract versnipperd heeft kan de records manager je vertellen wanneer dat precies gebeurt is en met welke reden.

¹⁵ All Change, Rich Robinson, 100 Movements Publishing (20 Sept. 2024).

Een record kan van alles zijn, zowel fysiek als digitaal. Dit boek is bijvoorbeeld een record. Wij zijn gevraagd dit op te sturen naar de Koninklijke Bibliotheek. Zij zijn nu de records manager. Ook al zijn wij de auteurs en eigenaars, het is iemand anders die het bewaart en bewaakt. Waarschijnlijk wordt het opgeslagen in een daarvoor speciaal gekoelde ruimte. Ook Emails video's, DVD's en noem maar op kunnen tot record worden "gedeclareerd". Declareren, zo noemen ze dat nu eenmaal. Dat is dus een andere betekenis dan wanneer je je onkosten declareert.

Er is nogal eens verwarring tussen de begrippen "Records Management" en "Retention Management". Hieronder noemen we een aantal karakteristieken van de twee begrippen.

Records Management	Retention Management
Bewaakt de levenscyclus van een object vanaf het moment dat het object als record gedeclareerd is.	Gaat vooral over het moment dat een object wordt weggegooid.
Zorgt ervoor dat er onweerlegbaar bewijs van het beheer van het object kan worden overhandigd	Zit veelal in een standaard content management systeem ingebouwd en vooral bedoeld om informatie niet tot in sint juttemus te bewaren
Kent processen zoals "legal hold" om verwijdering te voorkomen	
Weg = weg	
Records kunnen niet zomaar verwijderd worden	De document eigenaar kan veelal zelf het document weggooien
Complexe bewaarregels	Eenvoudige bewaarregels
Federated Records Management = Record A is afhankelijk van Record B	

In de praktijk is de scheiding niet zo expliciet, maar het wordt uit bovenstaande wel duidelijk dat Records Management een stapje verder gaat dan Retention Management. Retention Management is duidelijk wel een belangrijk onderdeel van Records Management. Het weggooien gebeurt echter op een zeer gecontroleerde manier.

Wat zou de reden kunnen zijn dat we bepaalde informatie zo zorgvuldig moeten bewaren. Waarom is het zo belangrijk dat er niet gerommeld (“tampered”) kan worden met bepaalde informatie. Waarom moeten we kunnen bewijzen dat we de informatie tijdig hebben weggegooid? Hieronder een aantal drijfveren voor het toepassen van Records Management op jouw informatie.

Records Management drijfveren	Voorbeelden
Wet en regelgeving	Persoonsgegevens moeten volgens de GDPR na verloop van tijd worden verwijderd.
Fiscale wetgeving	De boekhouding moet minimaal 7 jaar bewaard blijven om aan te kunnen tonen dat een bedrijf aan de fiscale wetgeving heeft voldaan.
Business Requirements	Een research bedrijf wil mogelijk een uitvindingen vastleggen ook als deze nog niet direct toepasbaar is. De documentatie mogelijk in combinatie met de fysieke resultaten worden als record gedeclareerd.

Records Management drijfveren	Voorbeelden
Administratieve noodzaak	<p>Je hebt een artikel gekocht met twee jaar garantie. Je moet wel dat bonnetje bewaren! Heb je het wel ondergebracht bij je records manager? De kans is groot dat je het na een paar maanden kwijt bent, of erger, het is verbleekt, onleesbaar!</p> <p>Ik ben ook heel blij dat de salarisadministratie mijn gegevens goed bewaard, alhoewel, ze weten het ook precies als ze teveel betaald hebben.</p>
Historische noodzaak	<p>Denk hierbij aan kadastrale gegevens. En zoals gezegd ook dit boek zodat onze slimme ideeën bewaard blijven voor het nageslacht.</p>

Dit kunnen allemaal drijfveren zijn om Records Management in te zetten.

Wanneer er inzicht is in de belangrijke informatie en de eisen voor het beheren hiervan kan een zgn. "Fileplan" worden opgesteld. Een Fileplan is een hiërarchie en categorisering van alle "records".

Opdracht:

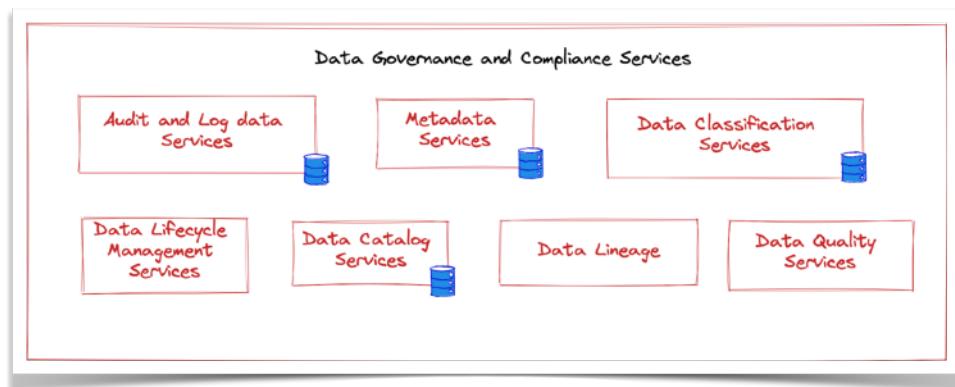
1. Wat is de belangrijkste informatie binnen jou organisatie die onder beheer van een Records manager valt of zou moeten vallen?
2. Is er in het bedrijf een Records Manager?
3. Welke policies zijn er binnen jouw bedrijf gedefinieerd?

We noemen dit categorieën. Aan deze categorie kun je dan weer je policies toekennen. Vaak zie je dat het hoogste niveau van het Fileplan de verschillende afdelingen zijn. Bijvoorbeeld HR (Administratieve noodzaak), Finance (Fiscale Wetgeving),

Procurement (Wet en Regelgeving). Je ziet dat in de basis policies van toepassing zijn op alle gegevens binnen de afdeling.

Op deze manier heb je een mooi overzicht gecreëerd van de informatie die voor jouw bedrijf van cruciaal belang is. Een Fileplan geeft structuur aan je informatie en die structuur is nodig om technologie in te zetten voor het managen van jouw records.

13. DATA GOVERNANCE AND COMPLIANCE SERVICES



Data Governance and Compliance Services

Data governance- en compliance services zijn van cruciaal belang omdat het inzicht geeft welke data beschikbaar is, waar die data zich bevindt en hoe en door wie deze kan worden gebruikt. Daarnaast verbeteren deze services de data kwaliteit en biedt het lineage om datastromen inzichtelijk te maken. Ten slotte ondersteunen ze de naleving van policies en wettelijke verplichtingen.

Even ter herinnering: technische governance is nooit genoeg; een organisatiecultuur heeft ook vaak de juiste veranderingen nodig en support van het management is hierbij ook noodzakelijk.

13.1. DATA GOVERNANCE VERSUS AI GOVERNANCE

Data- en AI governance zijn twee verschillende aandachtsgebieden die hand in hand gaan. Ze zijn allebei belangrijk en essentieel. Als je kijkt naar data governance dan is dit bedoeld om onze *data* onder controle te houden. We zorgen ervoor dat de data alleen nog maar toegankelijk is middels oplossingen als data catalogs. In een data catalog wordt allerlei informatie over data objecten bijgehouden. Door de snelle opkomst van AI met machine learning (ML), analytische toepassingen, Generative AI, etc. is het noodzakelijk geworden om meer specifieke governance te introduceren. Er zijn tools beschikbaar die iedereen in staat stellen (professional of niet) om AI te gebruiken (bijv. met Low-code no-code). Dit overstijgt de grenzen van data governance en start de wereld van *AI/governance*.¹⁶

Data governance heeft focus op beschikbaarheid, toepasbaarheid, integriteit en veiligheid van data. Data moet accuraat en consistent zijn en het moet verantwoord gebruikt worden conform wet- en regelgeving en bedrijfspolicies.

AI governance betreft de processen en policies met betrekking tot de ontwikkeling en de uitrol van AI projecten. Het dwingt projecten om zich te houden aan regels, processen en requirements voor AI initiatieven. Denk hierbij aan documentatie, risk management, bias, fairness, auditability en accountability van AI systemen.

De EU AI Act geeft aan welke governance maatregelen genomen moeten worden op basis van het risico dat de toepassing van het AI model met zich mee neemt. Voldoe je niet aan de AI act dan hangt er een behoorlijke zware boete boven je hoofd!

¹⁶ <https://blog.dataiku.com/spot-the-differences-data-governance-ai-governance#:~:text=Data%20governance%20ensures%20the%20accuracy,societal%20values%20and%20organizational%20priorities.>

Hoe implementeer je data governance en AI governance in je operatie?

Data governance implementeer je door:

- het implementeren van data policies,
- data toegankelijk te maken middels gecentraliseerde data catalogs,
- inrichten functioneel beheer van data (bijvoorbeeld data stewardship),
- het toepassen van data quality management processen.

AI governance implementeer je door:

- het implementeren van ethische policies,
- het toepassen van risk assessment raamwerken,
- het monitoring op bijvoorbeeld model drift.

Terwijl data governance in onze blauwdruk is ondergebracht in het architectuur domein Data Governance en Compliance Services is AI governance een onderdeel van het domein Data Analysis Services. In het ModelOps framework zitten de services die worden gebruikt voor AI governance.

Waarom is AI governance belangrijk?

Een AI governance framework is iets anders dan het in kaart brengen en catalogiseren van datasets. Een AI governance framework zorgt voor een strak maatpak als het gaat om bepaalde regels. Dit is absoluut noodzakelijk om fouten te voorkomen in de ontwikkelfase en het operationaliseren van AI. Er zijn nu eenmaal strikte regels over de inzet van AI (AI act). Generative AI (op basis van large language models) heeft ook governance nodig om langere termijn consequenties te voorkomen zoals data exposure¹⁷, hallucination effects¹⁸ en toxic

¹⁷ <https://www.infosecurity-magazine.com/blogs/accidental-data-exposure-gen-ai/>

¹⁸ [https://en.wikipedia.org/wiki/Hallucination_\(artificial_intelligence\)](https://en.wikipedia.org/wiki/Hallucination_(artificial_intelligence))

outcomes¹⁹ (vergeef ons deze Engelse termen maar kijk naar de literatuur die hierover is gepubliceerd).

Data op orde!

Een belangrijke standaard voor het inrichten van een goede structuur voor data is het DMBoK raamwerk.

DAMA-DMBoK (Data Management Body of Knowledge) dient als een raamwerk voor het begrijpen en implementeren van effectieve datamanagement praktijken. Nu data steeds belangrijker wordt voor de bedrijfsvoering, bijvoorbeeld in de banksector en de financiële dienstverlening, is een gestructureerde aanpak voor het beheren van data essentieel. Dit hoofdstuk biedt een samenvatting van de DAMA-DMBoK, met de nadruk op belangrijke gebieden zoals data governance, datakwaliteit en data-afkomst, met specifieke voorbeelden en toepassingen in de bancaire en financiële dienstverlening.

Wat beschrijft de DMBoK?²⁰

De DMBoK omvat processen, best-practices en principes op het gebied van Data Management.

Het richt zich op 11 hoofdonderwerpen:

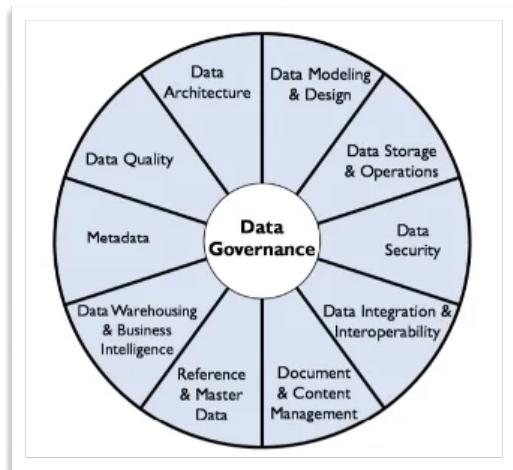
1. Data governance
2. Data architectuur
3. Gegevensmodellering en -ontwerp
4. Gegevensopslag en -bewerkingen
5. Data beveiliging
6. Gegevensintegratie en interoperabiliteit
7. Documenten en inhoud
8. Referentie- en master data
9. Data-warehousing en bedrijfsinformatie
10. Metadata

¹⁹ <https://www.linkedin.com/pulse/toxic-ai-prof-ahmed-banafa-38twc>

²⁰ <https://www.dataversity.net/what-is-the-data-management-body-of-knowledge-dmbok/>

11. Data kwaliteit

Daarnaast bevat de DMBoK speciale secties over data-ethiek, big data, data-integratie en -interoperabiliteit, en een maturity assessment van datamanagement. Om te illustreren hoe deze verschillende onderwerpen met elkaar verbonden zijn, heeft DAMA International het Data Management Framework of DAMA Wheel gecreëerd. Het omvat de 11 kennisgebieden en stelt Data Governance centraal, om te benadrukken hoe het ten grondslag ligt aan elke Data Management-activiteit.



DAMA Wheel

Doelstellingen van de DMBoK:

1. informatie verstrekken over best practices, rollen en verantwoordelijkheden, resultaten en statistieken, en maturity modellen voor gegevensbeheer
2. data managementpraktijken in het hele veld standaardiseren
3. te dienen als een functioneel raamwerk voor de implementatie van deze praktijken
4. een formeel vocabulaire voor deze concepten opstellen
5. de reikwijdte verduidelijken van wat deze praktijken wel en niet kunnen doen
6. leveranciers onafhankelijk overzicht bieden van managementpraktijken en mogelijke alternatieven voor specifieke situaties

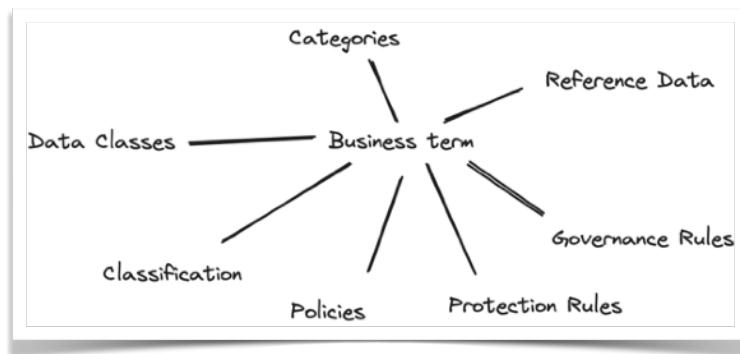
Een enorm handig raamwerk om je data management processen in te richten!

13.2. EEN DATA CATALOGUS, HOE MOEILIJK KAN HET ZIJN?

Tijdens discussies die plaatsvonden nadat we de eerste en tweede druk van dit boek hadden uitgebracht, kwam het onderwerp van een datacatalogus meerdere keren ter sprake. Je denkt misschien dat we de datacatalogus services impliciet hebben gedefinieerd als één groep centraal beheerde catalogus servicecomponenten die alle catalogusfunctionaliteit over alle soorten data biedt. In de ideale wereld zou dat het geval zijn. Een centrale opslagplaats die alle zakelijke termen, classificaties, categorieën, dataklassen, beleid en wat je nog meer nodig hebt om de gegevens voor je hele bedrijf te identificeren en te beschrijven.

Zowel voor al je gestructureerde als voor alle ongestructureerde data.

Laten we eens even kijken naar de gegevens die in een catalogus kunnen worden vastgelegd. De onderstaande afbeelding toont een overzicht. We hebben de business term als centrale definitie genomen en zoals je kunt zien heeft deze betrekking op verschillende objecten. Deze objecten hebben ook een relatie met elkaar, maar die zijn hier, om het simpel te houden, niet weergegeven.



Business term. Enterprise-vocabulaire om de betekenis van gegevens te beschrijven.

Categories. Een logische structuur om alle artefacten hiërarchisch te ordenen (betekenis, relatie).

Data classes. Classificatie van gegevens op basis van de structuur, het formaat en het bereik en waarde van de gegevens.

Reference data. Logische groeperingen van codewaarden (referentie gegevens waarden), zoals productcodes.

Classification. Beschrijf specifieke kenmerken van de betekenis van datagevoeligheid (bedrijfstermen en categorieën).

Policies. Beleid dat beschrijft hoe gegevens als assets moeten worden beheerd en beschermd. (Een beleid biedt je nog geen handhaving).

Governance rules. Beschrijving van hoe je een beleid toepast. Een natuurlijke taal van de criteria die moeten worden toegepast op gegevens assets. Handhaving dus!

Data protection rules. Er zijn regels geïmplementeerd die de toegang tot gegevens regelen op basis van gebruikers- en activa-eigenschappen en toegewezen governance-artefacten.

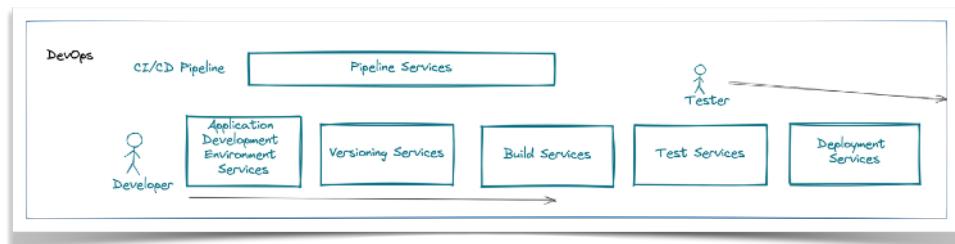
Je kunt erover twisten of het beleid moet worden afgedwongen door de catalogus of door de Data Security Services. Onze visie is hier eenvoudig. Het beleid zelf kan het beste in de catalogus worden opgeslagen, maar de handhaving ligt bij de Data Security Services. Maar dit is theorie en zal alleen werken als de twee naadloos integreren. Hiervoor zijn open standaarden nodig die nog niet zijn gedefinieerd.

Data Lineage Services	Data Lineage Services zijn diensten die de data stromen in kaart brengen. Het brengt de data stroom in kaart vanaf de bron tot uiteindelijke toepassing zoals een rapport of analyse. Het maakt inzichtelijk welke bewerkingen en door wie er op die data zijn uitgevoerd.
-----------------------	--

Data Catalog Services	De Data Catalog Services (data catalogus) levert de definitie van alle attributen en entiteiten van de data, dit kan op gebruksniveau of op technisch niveau zijn. Het verschaft de relatie tussen deze definities en bevat ook gegevens over de gebruikers van attributen en entiteiten.
Data Quality Services	Data Quality Services detecteert afwijkingen, waardoor je de kwaliteit van je gegevens kunt bepalen en de kwaliteit van je gegevens kunt verbeteren. Deze diensten worden ook wel data profiling genoemd.
Audit and Log Data Services	In een data-gedreven omgeving zullen beslissingen worden genomen door mensen en machines op basis van data en algoritmen. Audit gegevens leggen uit op welke basis beslissingen zijn genomen. Het bevat ook gegevens over wie of wat de beslissingen heeft genomen en zorgt voor volledige transparantie in de processen. In de loggegevens wordt beschreven wie wat heeft gedaan. In dit boek hebben we enkele woorden besteed aan het belang van AI-ethiek en compliance.
Metadata Services	Metadata Services biedt definities van ontologieën en taxonomieën met al hun relaties. Het is de representatie van de organisatie in een data structuur. Waar de data catalogus de data zelf beschrijft, beschrijft metadata hoe de entiteiten in de catalogus zich verhouden.

Data Life Cycle Management	<p>Records Management is een voorbeeld van Data Lifecycle Management voor ongestructureerde data. Soortgelijke diensten zijn echter ook van toepassing op gestructureerde data. Deze services worden ook wel data-archivering services genoemd, omdat in veel gevallen het gaat om het archiveren van oude data. De oude data moet verwijderd worden om de performance van je operationele omgevingen te behouden. Hoewel dit onderdeel kan zijn van jouw Systems of Record, wordt het resulterende archief hoogstwaarschijnlijk onderdeel van de “Systems of Insight”. Je data-analist bedankt je voor het verlenen van toegang tot deze data!</p>
Data Classification Services	<p>Data Classification Services wijst een dataklasse toe. Dit helpt bij het vastleggen van de gegevens die in data objecten staan. Dit is zowel van toepassing op gestructureerde als ongestructureerde data en kan vooraf worden bekijken door catalogus gebruikers</p>

14. CI/CD PIPELINE SERVICES



CI/CD Pipeline Services

De ‘continuous integration and continuous deployment’ oftewel de CI/CD-pipeline is belangrijk om onze applicaties op een agile manier te kunnen ontwikkelen en inzetten. Er is zoveel informatie beschikbaar over dit onderwerp, dat we alleen de belangrijkste services hebben vermeld. Er zijn nog veel meer services die je in je pipeline kunt opnemen.

Voordat we op de services ingaan een paar belangrijke afwegingen om te maken:

- Zorg ervoor dat jouw pipeline cloud onafhankelijk is.
- Forceer jezelf niet om voor één pipeline te gaan, maar ontwikkel een beperkt aantal pipelines voor verschillende type systemen. Wij noemen dat ‘fit for purpose’. Een CI/CD-pipeline die je bijvoorbeeld voor je Systems of Engagement hebt ontwikkeld, verschilt waarschijnlijk van de pipeline die je voor je Systems of Record gebruikt.
- Overweeg het gebruik van lichtere in de cloud geboren pipelines.
- Overweeg toekomstige ontwikkelingen die door jouw pipeline moeten worden ondersteund, zoals serverless.

- Zorg ervoor dat de pipeline veilig is. *DevSecOps* wordt als concept naar voren gebracht om ervoor te zorgen dat er geen kwetsbaarheden in je code zitten die je in productie brengt en om kostbare herontwerpen te voorkomen. Test minimaal OWASP kwetsbaarheden²¹.
- Neem duurzaamheid aspecten mee: hoe efficiënt is de pipeline zelf en de code die deze produceert? Hiervoor kun je *SusDevSecOps* introduceren. Naast secure by design ook sustainable by design!
- GitOps is in opkomst als concept. Overweeg dit concept omdat het je dwingt op een cloud native manier²² te denken en te werken waarbij de pipeline maximaal geautomatiseerd wordt.

We hebben de automatiserings- en provisioning diensten gepositioneerd in de Hybrid Cloud Management Services domein. Deze services kunnen worden aangeroepen vanuit de CI/CD Pipeline.

Pipeline Service	De CI/CD-pipeline service is de orkestratie laag die specifieke services met elkaar verbindt en die samen een end-to-end-proces creëren, van het maken van code tot de implementatie van een service in productie. Het aantal beschikbare diensten voor het samenstellen van de pipeline is extreem groot. XebiaLabs creëerde een mooie kaart (periodiek systeem genoemd) van 'alle' diensten gerelateerd aan hun doelstellingen (testen, versiebeheer, implementeren, ...). In ons geval vermelden we alleen de belangrijkste. Als je met nieuwe ontwikkeling begint, is het de moeite waard om naar tools als Tekton en GraalVM te kijken in plaats van naar het goede oude Jenkins.
------------------	--

²¹ <https://owasp.org/www-community/vulnerabilities/>

²² <http://ibm.biz/ProgModel>

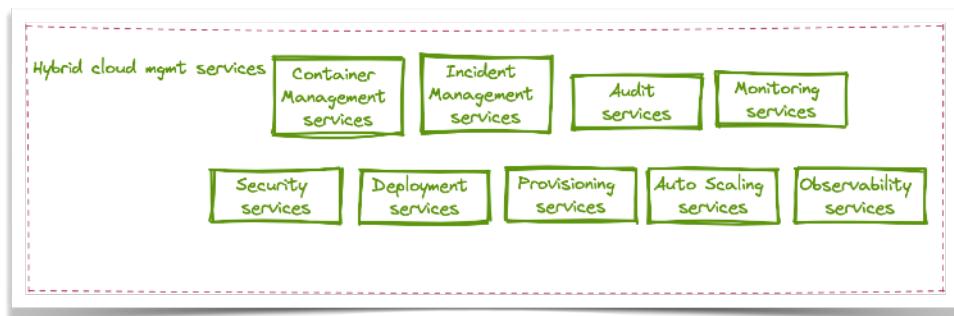
Application Development Environment (ADE) Services	<p>De applicatie ontwikkel omgeving is de omgeving waarin een applicatieontwikkelaar een unittest kan bouwen en uitvoeren. Heel belangrijk in deze omgeving is dat ontwikkelaars snel kunnen schakelen tussen het wijzigen van code en het uitvoeren van tests. Code schrijven en testen is een continue cyclus die vele malen per dag plaatsvindt. Performance is de belangrijkste succesfactor voor deze omgeving. In een data-gedreven organisatie ontwerpen en bouwen ontwikkelaars vooral diensten die toegang bieden tot data, datamodellen, algoritmen en bedrijfsregels om het bedrijfsproces te verbeteren.</p>
Versioning Services	<p>Versiebeheer kan zeer complex zijn in een omgeving waarin veel releases (versies) van een applicatie draaien. Versiebeheer moet worden geïntegreerd in de CI/CD-pipeline. Zodra versies zijn "ingecheckt" in het systeem, begint de CI/CD-pipeline met het bouwen, testen en implementeren van een nieuwe versie van de code. Een voorbeeld is Git, het wordt een trend om alles vanaf daar te laten draaien.</p>
Build Services	<p>De build services brengen code libraries en code samen in een compiler om een applicatie (microservice) te creëren die in de operationele omgeving kan draaien.</p>
Test Services	<p>Testservices zijn die services die allerlei tests kunnen uitvoeren op de ontwikkelde applicatie. Bij DevSecOps zijn beveiligingstests van cruciaal belang.</p>

Deployment Services	In de data-gedreven omgeving worden applicaties in de veel gevallen deployed in een container runtime. Meestal wordt die container omgeving beheerd door Kubernetes. Kubernetes verzorgt een groot deel van de implementatie activiteiten, het wijst bijvoorbeeld de runtime toe, wijst schijfruimte toe en installeert monitoring tools, enz.
---------------------	--

14.1. DEVOPS VERSUS MLOPS

Er is een verschil en een overeenkomst. Beide zijn software ontwikkel strategieën voor het ontwikkelen, testen en operationaliseren van toepassingen. Het verschil is dat DevOps zich focust op het ontwikkelen van applicaties terwijl MLOps zich richt op Machine Learning. Zie het hoofdstuk frameworks voor een verdere verdieping.

15. HYBRID CLOUD MANAGEMENT SERVICES



Hybrid cloud management services

Een data-gedreven organisatie heeft de mogelijkheid om datadiensten als microservices beschikbaar te stellen en deze op een containerplatform te draaien. Deze containeromgeving moet beheerd worden. Door de grote hoeveelheid microservices en containers die beheerd moeten worden, wordt het voor de mens vrijwel onmogelijk om de hele omgeving met honderden containers en duizenden verbindingen in de gaten te houden. Daarom is AIOps ontstaan. AIOps vereist een breed scala aan tools. De hybride cloud management services biedt een set tools voor monitoring, observatie en systeembeheer. AI wordt gebruikt om incidenten te detecteren en zelfs te voorspellen en doet voorstellen om incidenten op te lossen. Door de grote hoeveelheid services is er het risico dat omgevingen gemakkelijk overgedimensioneerd worden. Met de nadruk op duurzaamheid en kosten efficiency is overmatige toewijzing van CPU en storage niet langer acceptabel en applicatie resource management zorgt ervoor dat de omgeving de juiste capaciteit levert.

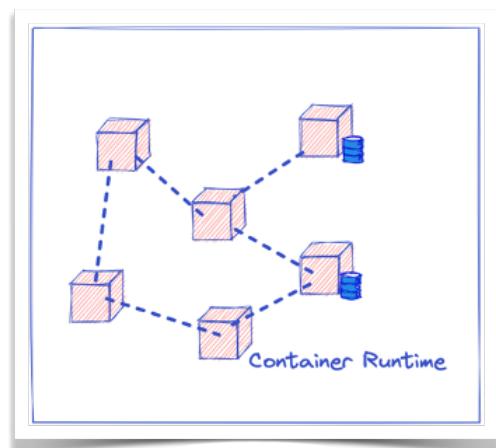
Container Management Services	Kubernetes is de standaard beheer service voor het beheren van containers. Het biedt schaalbaarheid en zelfherstellende mogelijkheden.
Incident Management Services	Registreert het incident en monitort de acties om het incident op te lossen vanaf het begin totdat het incident ook daadwerkelijk is opgelost. De Incident Management Services maakt deel uit van de Systems of Record, maar zoals we hebben gezien, worden de Systems of Insight in een data-gedreven organisatie steeds meer onderdeel van de Systems of Record. Op het moment dat je de Sol gebruikt om de SoR te controleren, wordt deze er in wezen deel van en ligt het voor de hand om gebruik te maken van dezelfde services.
Audit Services	Audit Services verzamelt lineage- en audit gegevens en je kunt deze service gebruiken voor het maken van rapporten ten behoeve van compliance.
Monitoring Services	Monitoring Services rapporteert voortdurend de status van hardware, besturingssysteem, middleware, netwerk, virtualisatie en applicaties en signaleert (potentiële) problemen op basis van drempelwaarden zoals CPU-, geheugen- en opslaggebruik.

Security Services	Security Services is een verzameling van allerlei beveiligingsdiensten voor identiteitsbeheer, privileged access management, vulnerability testen, zero trust, malware, inbraakdetectie etc. Security Services zijn niet in detail uitgewerkt omdat ze op alle systemen van toepassing zijn. In een data-gedreven organisatie moet speciale aandacht uitgaan naar <i>databeveiliging</i> , het meest waardevolle bezit in jouw organisatie. Maak het <i>secure by design</i> .
Deployment Services	De Deployment Services zijn functioneel dezelfde als de services vanuit de CI/CD-pipeline. Maar niet alle applicaties maken deel uit van een pipeline, bijvoorbeeld de applicatie die je hebt gekocht, middleware-producten, enz. Architectuur beslissingen bepalen welke deployment services je gebruikt.
Provisioning Services	Deze dienst vormt de kern van ‘Infrastructure as Code’. Het heeft de mogelijkheid om infrastructuur te creëren waar de applicaties op kunnen draaien. Je voert bijvoorbeeld een script uit om een nieuw vLAN te creëren of je bouwt een script om storage toe te wijzen. Er zijn verschillende scriptingtools zoals Chef, Puppet en Terraform, maar de meest algemeen geaccepteerde is Ansible. Cloudproviders hebben vaak weer hun eigen tools. Het is een fundamentele architectuur beslissing om voor een generiek tool of voor een cloudspecifiek tool te kiezen. Onze multi-hybride cloudvisie stemt duidelijk voor de eerste optie.

Auto Scaling Services	Enerzijds is dit de mogelijkheid om te schalen binnen de containeromgeving maar ook in de traditionele omgeving, bijvoorbeeld VM's. Kubernetes kan worker nodes en machines dynamisch aan clusters toewijzen. In de VM-wereld zijn er diensten beschikbaar om op- en af te schalen. Automatisch schalen wordt een belangrijke duurzaamheidsvereiste. Tevens is dit een belangrijke capability om je kosten in de cloud te bewaken.
Observability Services	Geeft inzicht in de infrastructuur en de relaties binnen de infrastructuur. Het bepaalt de performance van jouw systemen en kan worden gebruikt om de hoofdoorzaken van problemen te vinden en te analyseren. Dit wordt ook wel applicatie performance monitoring (APM) genoemd.

15.1. APPLICATION RUNTIME

Het klinkt misschien een beetje voor de hand liggend. Je hebt een applicatie-runtime nodig om je applicaties te draaien. En waarom zou dit niet een microservices-architectuur zijn die op containers draait? In februari 2022 publiceerde Randy Bean een artikel in Harvard Business Review. Als belangrijke oplossing voor het overwinnen van de obstakels luidt zijn advies: "Fail fast, learn fast". Een op microservices gebaseerde architectuur ondersteunt dit advies. Een



Container runtime

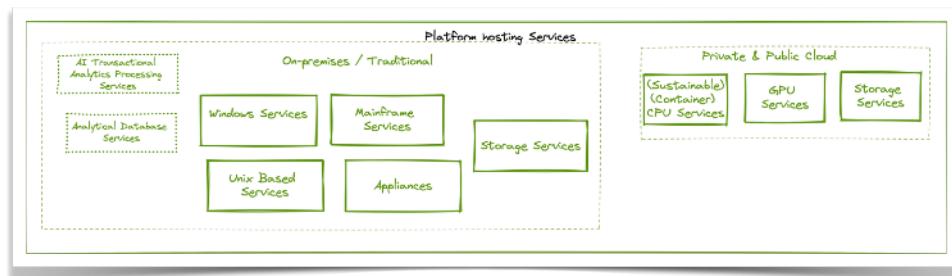
tweede reden waarom deze runtime erg belangrijk is, is dat deze overal moet kunnen draaien.

De applicaties moeten dezelfde runtime beschikbaar hebben als ze draaien in de private cloud, de publieke cloud of als ze draaien op maatwerk technologie, maar ook als deze draait op moderne RISC gebaseerde duurzame technologie. Het zou zelfs als edge technologie moeten kunnen draaien indien er een noodzaak is om gegevens aan de edge te verwerken. Bijvoorbeeld om hoge kosten voor uitgaand verkeer (egress) te vermijden of als er simpelweg geen bandbreedte beschikbaar is.

Wij zijn van mening dat de keuze voor een runtime-omgeving (bijvoorbeeld AKS, OpenShift, GKE) een belangrijke beslissing vraagt die weloverwogen gemaakt moet worden. De runtime-omgeving is een essentieel onderdeel van de blauwdruk voor de enterprise referentie architectuur.

De keuze van de runtime kan verband houden met de CI/CD-pipeline. De combinatie van de runtime-omgeving en de CI/CD-pipeline moet het proces van ‘Fail fast, learn fast’ ondersteunen. Later zullen we het belang van ModelOps benadrukken om dezelfde snelheid te bieden bij het maken van AI modellen.

16. PLATFORM HOSTING SERVICES



Platform Hosting Services

Het idee is dat de infrastructuur is gebaseerd op cloud concepten. Denk hierbij aan schaalbaarheid, doorbelasting, flexibiliteit, etc. Niet noodzakelijkerwijs een publieke cloud, het zou heel goed een private cloud kunnen zijn. Voor data-gedreven organisaties is het erg belangrijk om rekening te houden met de potentiële kosten van egress en ingress voordat je beslist waar je data analyses gaat uitvoeren. Over het algemeen kun je de analyses het beste uitvoeren op de locatie waar je gegevens zijn opgeslagen.

Onze stelling is: voor iedere type werklast zoek je het juiste platform om dit te hosten. Zijn het missie kritische werklasten met een zeer hoog transactievolume en een groot aantal gebruikers die gelijktijdig op het systeem werken dan zou je mainframe services kunnen overwegen. Veel applicaties worden ontwikkeld voor interactie met eindgebruikers zoals portal services of andere apps die gebruikers in staat stellen om informatie op te vragen of informatie op te sturen. Deze informatie kan bijvoorbeeld worden gebruikt als input om een proces in een andere business toepassing te starten. Dit soort systemen worden meestal op

'distributed systemen' op basis van Windows of Linux ontwikkeld. Meestal worden zelfgebouwde applicaties ontwikkeld op Linux. Open Source is toch wel een belangrijke trend in de markt.

Op wat voor systeem je applicaties dan ook gaat hosten: in alle gevallen moet je nadenken over hoe je de bedrijfsvoering kan garanderen in het geval van calamiteiten. Hiervoor hanteren bedrijven 'disaster recovery' policies en procedures. Het gaat hierbij niet alleen om de beschikbaarheid van digitale systemen en de data. Het is ook belangrijk om te kijken wat je wel kan doen als deze systemen tijdelijk of helemaal niet beschikbaar zijn. Wat kan het bedrijf doen om een dienstverlening in afgeslakte vorm zonder systemen uit te laten voeren? Papiertjes, telefoonjes, lijst met backlog bijhouden om die later weer te 'verzoenen' (reconcile in het Engels) in de systemen..... Allemaal een onderdeel van een business continuity en disaster recovery plan. Gelukkig hebben de systemen zelf een groot technisch vermogen om bij storingen of uitval (bijv. stroomstoring) door te blijven draaien. Veel technische concepten voorzien hier vandaag de dag in. Denk aan replicatie van data over twee locaties, of clusters van servers die bij gedeeltelijke uitval kunnen doordraaien, netwerkpaden die redundant zijn uitgevoerd, etc.

Containerisatie van cloud native applicaties introduceert een nieuwe vorm van hoge beschikbaarheid. De infrastructuur is dan niet langer de belangrijkste factor in beschikbaarheid maar de flexibiliteit van container architectuur levert de beschikbaarheid. Bij het falen van de infrastructuur is de container service in staat om de applicatie te herstarten op een ander infrastructuur platform die binnen het domein van de container service draait. Let op, dit moet wel een hardware platform zijn van dezelfde architectuur omdat bij verschillende hardware architecturen de container (applicatie) niet binair compatibel is. Met andere woorden, de applicatie is in de build fase gelinkt aan het onderliggende hardware platform. Deze applicatie kan dus alleen

op deze specifieke hardware architectuur draaien. Wil je de applicatie op een ander hardware platform draaien dan zullen je die moeten ‘bouwen’ (build) op dit specifieke platform! Dit is het concept rond immutable infrastructuur die je ook kan inrichten op basis van containers.

16.1. ANALYTICAL WERKLASTEN

Het tweede onderwerp om te overwegen zijn de analytische infrastructurele mogelijkheden, met andere woorden, de GPU-processors. Als je veel werk hebt voor deze type processoren, wil je deze als ontwikkelaar graag ‘claimen’.

Maar.... dat willen ze allemaal!

Bij grote werkdruk kunnen eventueel GPU’s vanuit de cloud worden ingezet.

Ten derde de vraag of analytische capabilities direct op de operationele data nodig zijn. Een voorbeeld hiervan is een creditcard transactie. Om de transactie te valideren aan de hand van analytische modellen, heb je deze capabilities op je operationele systemen nodig. Vanuit het perspectief van een data-gedreven organisatie zijn wij van mening dat dit voor jouw bedrijf een onderscheidende factor zou kunnen zijn om je transacties beter te beveiligen en te valideren terwijl ze worden uitgevoerd.

16.2. HOSTING DIENSTEN IN CONTEXT VAN ANALYTICS

We hebben de platform hosting services opgesplitst in drie soorten hosting diensten:

1. De private cloud. Veel bedrijven gebruiken dit als productieomgeving om de modellen uit te voeren. Waar modellen worden getraind, is afhankelijk van de hoeveelheden en locatie van de data. De Learning Accelerator Service is van onbetwiste waarde voor het uitvoeren van de planningen voor

modeltrainingen. Deze service kan ervoor zorgen dat de kostbare GPU processor voor 95% van de tijd gebruikt wordt.

2. De publieke cloud. Deze zou gebruikt kunnen worden als ontwikkel- en productie omgeving. Specifiek voor kleinere bedrijven die een beperkt aantal modellen hebben om te trainen of niet over de schaalgrootte en mensen beschikken om hun eigen omgeving te beheren. Als je modellen gaan bouwen en je toepassingen naar verwachting zullen gaan groeien, is het wel raadzaam om tools te gebruiken die niet alleen beschikbaar zijn in de publieke maar ook in de private cloud om cloud lock-in te voorkomen.
3. De bestaande on premise omgeving waar je mogelijkheden beschikbaar stelt voor real-time integratie van de Systems of Record en Systems of Insight. Omdat de oplossingen dicht bij de processor hardware en/of het operationele database management systeem worden ontwikkeld, zijn deze toepassingen heel specifiek en soms proprietary van aard.

AI Transactional Analytics Processing Services	Een platformservice voor het uitvoeren van real time AI analyses op transacties. Om dit te doen bevat de processor die de transactie uitvoert een GPU. Er zijn verschillende processoren op de markt die deze dienst ondersteunen.
Business Continuity Services	Data-gedreven organisaties worden steeds afhankelijker van de 'Systemen'. Ook in de context van Sol is de behoefte aan kwalitatieve aspecten zoals hoge beschikbaarheid en disaster recovery gegroeid. Verschillende architectuur typologieën vereisen verschillende HA/DR-services.
Windows Services	Voor traditionele (Windows-gebaseerde) data-gedreven workloads die nog niet in containers kunnen draaien.

Unix Based Services	Voor traditionele, op Unix gebaseerde workloads, zoals Linux of gebaseerd op eigen besturingssystemen zoals AIX en HPUX.
Mainframe Services	Enterprise mainframe platform voor hoge transactionele werklasten. De focus ligt op de <i>real-time analytische mogelijkheden</i> van dit systeem dat vooral voor de Systems of Record wordt ingezet.
Database Accelerator	Met de Database Accelerator kun je analytische queries uitvoeren op een online operationele database. In de meeste gevallen wordt dit niet als een best practice gezien, maar als er op enig moment daadwerkelijke inzichten moeten worden geleverd op de actuele data is dit de oplossing. Gelukkig zijn er oplossingen beschikbaar die niet of nauwelijks invloed hebben op de performance van de operationele database.
GPU Services	GPU-diensten werken heel goed met neurale netwerken en zijn daarom ideaal voor AI. De GPU-services kunnen “stand-alone” GPU-services zijn of kunnen worden gecombineerd met een reguliere (CPU-)server. GPU-processoren zijn duur!. De Learning Accelerator Services kunnen de werklast optimaliseren waardoor de GPU optimaal wordt benut.

(Sustainable) (Container) CPU Services	Op RISC of CICS gebaseerde processors voor het runnen van applicaties. Deze verwerkingskracht kan lokaal of in de cloud beschikbaar zijn. Afhankelijk van hoe serieus je duurzaamheid neemt, kunnen er aanzienlijke hardware verbeteringen worden aangebracht. Op containers gebaseerde workloads kunnen efficiënter draaien als gebruik gemaakt wordt van een ARM, RISC-V, OpenPower of Telum gebaseerde processor architectuur.
Storage Services	We hebben data opslag gecombineerd in één eenvoudig blok met opslagdiensten. Je zult onderscheid moeten maken tussen blok-, file- en objectstorage, maar ook tussen actieve, passieve en archiefstorage en uiteindelijk back-up. Dit hebben wij uitgelegd in de volgende paragraaf over storage. Maar het mooie van een dienst is dat je hier voorlopig geen zorgen over hoeft te maken: deze storage diensten beheren alle data transparant voor jou, van (over)provisioning tot storage tiering, het is allemaal geregeld.
Appliances	Er zijn veel apparaten op de markt die een hele specifieke functie uitvoeren. De appliance doet één ding en doet dat heel efficient. Van Teradata-platforms voor databases tot Data Power voor applicatie integratie. Als je veel van één functie gebruikt kan een appliance een ontlasting zijn voor je beheer organisatie.

16.3. STORAGE

In onze blauwdruk heb je één dienst “Storage services” gevonden. De systems of insight zijn groot data gebruikers, daarom is het vanuit dat perspectief de moeite waard om wat

woorden te besteden aan de opslag van die grote hoeveelheden data.

Allereerst moet je onderscheid maken tussen blok-, file- en object storage. Ten tweede moet je onderscheid maken tussen opslag media zoals Flash, Tape, Disks. Ten derde moet je nadenken over de toegangskenmerken zoals actief, passief en archief en ten slotte moet je nadenken over back-up en hoge beschikbaarheid. En om het nog ingewikkelder te maken: er is een relatie tussen die aspecten soms sluiten ze elkaar zelfs uit. Je gebruikt bijvoorbeeld geen objectstorage op tape voor actieve toegang.

Storage tiering is een concept waarmee je gegevens transparant tussen opslagmedia kunt verplaatsen. Andere diensten zijn: storage virtualisatie, storage compressie en over-provisioning.

Opdracht:

Breng je verschillende type storage in kaart en voor welke toepassing deze worden ingezet.

Je hebt misschien het gevoel dat gegevensopslag niet erg duur is, dus waarom zou je hier druk over maken?

En in zekere zin is dat juist. Opslag is erg goedkoop geworden als je 1 GByte of zelfs een paar terabytes opslaat, dan is dat geen probleem. En daar zit een factorje 1000 tussen.

Als je het nog eens vermenigvuldigt met 1000 wordt het 1 Petabyte en ineens wordt het wel een kostenfactor.

Vermenigvuldig nu alleen deze 1 PetaBytes met een factor 10 en het wordt een onderwerp in het managementoverleg en vermenigvuldig we dat weer met slechts een factor tien en het wordt een bestuurskamer gesprek.

Daarom moet je een opslagstrategie ontwikkelen, waar laat je al die gegevens?

16.4. HYPER-CONVERGED INFRASTRUCTUUR

Werkt hyper-converged infrastructuur in een data-gedreven organisatie?

Als je bedrijf een hyper-converged infrastructuur strategie heeft, is opslag, CPU en netwerk allemaal samengevoegd in één box (wat ideaal is voor Systems of Engagement). Maar je moet beoordelen of dit ook past bij data-intensieve gebruik scenario's.

16.5. QUANTUM ALS HOSTING PLATFORM

Even een blik in de toekomst.....Waar gebruiken we Quantum voor, een voorbeeld.

Quantum kan worden ingezet voor diverse toepassingen. Denk hierbij aan optimalisatie vraagstukken, zeer complexe wiskundige berekeningen, natuurkundige en biologische vraagstukken. Kortom in domeinen waar traditionele computers over struikelen of er eer lang over doen om met een resultaat te komen. Denk aan de ontwikkeling van medicijnen bijvoorbeeld.

Zoals de term ‘optimalisatie vraagstuk’ al doet vermoeden bestaat er geen eenvoudige formule waarmee een dergelijk probleem kan worden opgelost. De klassieke computer berekent eerst allerlei mogelijkheden om deze vervolgens te vergelijken en de beste eruit te halen. Naarmate er meer variabelen zijn, hebben we het in de volksmond al gauw over een onbeperkt aantal mogelijkheden. De klassieke computer analyseert alle mogelijkheden één voor één en stopt er na verloop van tijd ook mee onder het motto “goed is goed genoeg”. Een Quantum Computer is in staat om een grote hoeveelheid van deze berekeningen parallel uit te voeren. Dat betekent twee dingen: het gaat vele malen sneller en het kan ook meer berekeningen uitvoeren waardoor het resultaat nauwkeuriger wordt.

Ondanks de potentie van Quantum Computers zal het niet helemaal een fluitje van een cent zijn, maar zou het wel een stuk sneller en misschien ook beter kunnen. Want wellicht komen we

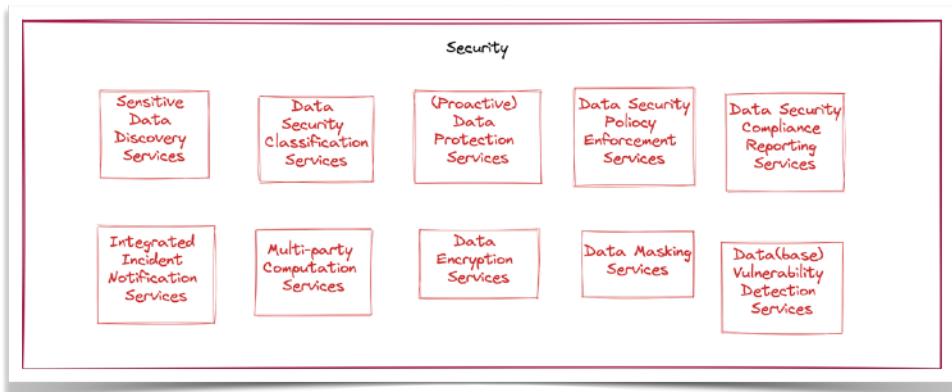
wel tot betere oplossingen/antwoorden doordat er verder doorgerekend kan worden. Daarmee hebben we niet alleen versnelling maar ook een verbetering van de kwaliteit van de uitkomst.

Dit neemt overigens niet weg dat er geen rol meer is voor de klassieke computer. Denk hierbij aan een hybride aanpak waarbij je een deel van de berekening op onze vertrouwde klassieke omgeving uitvoert en vervolgens de circuits van een Quantum Computer aanspreekt en het resultaat weer terugstuurt naar de ‘vertrouwde klassieke omgeving’ om hier verder mee te werken. Zou het niet mooi zijn als er een manier is om te bepalen wanneer dat moment aanbreekt, een soort van cut-off point (kantelpunt) op basis van bijvoorbeeld de hoeveelheid variabelen of de grootte van de dataset, et cetera? Dat zou betekenen dat je alleen bij bepaalde use-cases direct de Quantum Computer inzet, waar het duidelijk is dat het geen zin heeft om klassieke optimalisatie algoritmen in te zetten. Scheelt verspilling (het stampen van de computer zonder resultaat) en je komt met betere antwoorden.

Er zijn al stappen genomen om een klassiek optimalisatie algoritme te vergelijken met een Quantum algoritme. Binnen Qiskit (toolkit voor Quantum toepassingen) zijn er Quantum algoritmen beschikbaar die je zou kunnen vergelijken met bijvoorbeeld een klassiek Cplex algoritme (tool voor traditionele optimalisatie algoritmes).

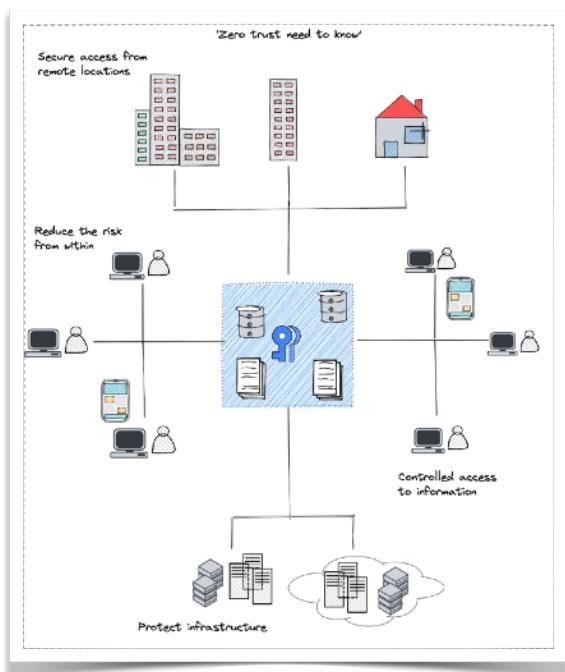
Dit biedt de mogelijkheid om benchmarks te kunnen uitvoeren om op korte termijn veel concreter te kunnen gaan testen en vergelijken.

17. DATA SECURITY SERVICES



Data Security Services

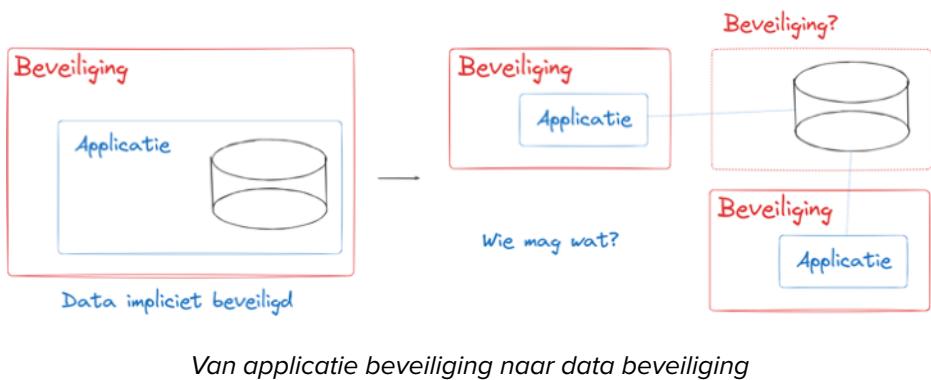
Vaak zie je dat data security geen integraal onderdeel is van een architectuur voor data-gedreven organisaties. Wij zijn echter van mening dat hoe meer organisaties met data als essentieel instrument werken, hoe meer deze tegen kwaadaardige invloeden moeten worden beschermd. Daarom is het onze mening dat data-gedreven organisaties



een security architectuur moeten adopteren die gebaseerd is op ‘Zero trust, need to know’-principes.

Een fundamentele fout die we ook hebben vastgesteld, is dat een goede cybersecurity strategie niet voldoende is om je gegevens te beschermen.

Zoals we allemaal weten is het gevaar en de gevolgen van aanvallen (het wijzigen, verwijderen, stelen of delen van gegevens) zeer ernstig. De noodzaak als gevolg van de Covid-pandemie om medewerkers vanaf elke locatie en vanaf elk apparaat te kunnen laten werken heeft de behoefte aan geavanceerde beveiligingsarchitecturen versneld.



Traditioneel werden applicaties gebouwd met toegangscontrole (wie mag deze applicatie gebruiken) en binnen de applicatie werden delegatieregels toegepast om controle te hebben over wie welke informatie mag zien/toegang krijgen of wie een proces mag uitvoeren. Omdat data in andere processen wordt gebruikt en op veel andere manieren wordt benaderd, zijn andere security oplossingen nodig om controle te krijgen over de toegang tot de bedrijfsgegevens.

Security, Compliance en Governance zijn drie onderwerpen die verschillend zijn, maar wel gerelateerd. Met governance zorgen we ervoor dat we de controle hebben, met security

implementeren we die controle en compliance levert ons het bewijs. Laten we het niet moeilijker maken dan dit.

Waarom hebben we Data Security Services opgenomen in een specifiek architectuur domein?

In onze blauwdruk hebben we bijvoorbeeld Data Discovery Services, Classification Services, Masking Services etc.

gedefinieerd. Deze zijn ook security capabilities maar die zijn onderdeel van andere logische architectuur domeinen.

Ook al zijn het functioneel gezien dezelfde diensten, de realiteit is dat de technologieën behoorlijk verschillend zijn.

Hoe goed zijn je algemene Data Discovery Services vergeleken met de specifieke Sensitive Data Discovery Services? Deze laatste is specifiek gebouwd voor het detecteren van gevoelige data, dus je zou verwachten dat deze diensten zeer geavanceerd zijn en erg goed zijn in het begrijpen van de gevoeligheid van je gegevens en bijvoorbeeld je gegevens kunnen toewijzen aan de CIA Rating. **CIA** staat voor **C**onfidentiality, **I**ntegrity en **A**vailability. In de ideale wereld zouden de Security Services bij voorkeur één geheel van diensten zijn, maar zoals je wellicht weet leven we (nog) niet in deze ideale wereld. Dus misschien heb je vandaag beide sets Discovery Services nodig. Je zult merken dat de door ons gedefinieerde Data Security Services altijd een niveau dieper gaan.

In plaats van vast te stellen 'welk systeem heeft toegang tot welke database', willen we meer weten en controleren: 'wie heeft toegang tot welke data,

Waar Security Services te vergelijken zijn met een uitsmijter, zijn sommige Data Security Services net undercoveragenten, ze lopen rond in jouw organisatie, altijd alert op de persoon die zich verdacht gedraagt.

P.S. Ze aarzelen niet om de uitsmijter te bellen om deze persoon te elimineren....!

hoe heeft hij/zij toegang tot deze data en in welke context worden deze data benaderd, enz..! In beide gevallen is het een access control service, maar het niveau van de controle mechanismen maakt het verschil.

We hebben deze soms overlappende reeks diensten gedefinieerd om je aandacht te vestigen op de verschillende aspecten van data beveiliging.

Stel jezelf bij elk van deze diensten de vraag: wordt deze goed genoeg aangepakt als dienst in een ander domein of is het effectiever om specifieke data security services te implementeren?

Uit een recent onderzoek is gebleken dat de kosten van datalekken stijgen.

En natuurlijk zal een datalek jouw bedrijf nooit overkomen... totdat..... het gebeurt.

'Penny wise pound foolish' is het spreekwoord om te onthouden.

17.1. STEL JE EENS VOOR....

IT-beveiliging is een beetje een vreemd domein: voor velen een weinige geliefd onderwerp (lastig), gehaat door velen (al die restricties!), maar het is uiteindelijk noodzakelijk voor iedereen. IT-beveiliging is de afgelopen dertig jaar een extreem complex domein geworden waar het voortbestaan van bedrijven afhankelijk van is. Het is bijna dagelijks nieuws dat hackers de bedrijfsvoering platleggen door DDOS-aanvallen of dat organisaties worden gekaapt met ransomware en dat er grote sommen geld moeten worden betaald om weer toegang te krijgen tot hun IT-middelen.

Gijzeling is een bekende activiteit waarbij mensen met geweld een voorwerp (meestal vliegtuigen of treinen) 'stelen' om 'iets' te eisen, en bij ontvoering worden meestal kinderen of vrouwen gevangengenomen om losgeld te krijgen voor hun vrijlating. Als

men de aanwijzingen van de ontvoerders niet direct opvolgt kan dit verschrikkelijke gevolgen kunnen hebben voor de kinderen of de vrouw in kwestie. Vaak wordt losgeld betaald en tegelijkertijd probeert de politie een val te zetten om de ontvoerders te pakken te krijgen. Altijd een risicovolle onderneming voor de ontvoerders omdat er vaak de kans bestaat dat de (slimme) politie ze toch te pakken krijgt en ze achter de tralies gooit (of erger). Veel films hebben een groot kassucces behaald met dit onderwerp.

Zo anders in onze IT-wereld, waar IT-kapers en IT-ontvoerders niet eens hun bed hoeven uit te komen om hun duistere praktijken uit te voeren. Ze zijn anoniem, ze zijn onzichtbaar en er zijn genoeg potentiële slachtoffers.

In dit geval is het slachtoffer de eigenaar van gegevens en niet een vader van een kind of de vrouw van een rijke magnaat. En organisaties hoeven niet rijk te zijn. Een solide business case voor het betalen van het losgeld is goed voorbereid. De hoogte van het gevraagde losgeld is gerelateerd aan de financiële draagkracht van het bedrijf en de waarde die men waarschijnlijk ervoor over heeft. Het losgeld is dus altijd op maat!

Op 18 september 2023 kwam het volgende bericht in het nieuws over cyberaanvallen op producenten voor alternatieven voor Russische gasproducten:

'Cyberaanvallen vormen een bedreiging voor de importterminals voor vloeibaar aardgas (LNG). Het hoofd van de Duitse inlichtingendienst BND Bruno Kahl waarschuwt hiervoor. De terminals maken LNG, dat wereldwijd een belangrijk alternatief is geworden voor Russisch aardgas, geschikt voor gasleidingen'.

Dit voorbeeld laat zien dat cyberaanvallen niet alleen worden gebruikt voor financieel win, maar ook worden gebruikt om de mondiale stabiliteit te verstören. Andere voorbeelden zijn

aanvallen gericht op het verkrijgen van intellectuele informatie in hightech industrieën of overheidsinformatie.

Wie herinnert zich de klassieke film "War Games" uit 1983 (dat is vier decennia en honderden IT-technologieën geleden) waarin een jonge computer whizzkid "per ongeluk" inbreekt in een militaire centrale computer die de volledige controle heeft over het Amerikaanse nucleairearsenaal? Hij denkt dat hij een spel speelt, maar in feite begint hij bijna de Derde Wereldoorlog. Hoe bizar om deze verhaallijn te lezen van de film die veertig jaar geleden een reële situatie van hedendaagse horrorscenario's voorspelde.



Hoe zit het met de kans dat jijzelf met fraude te maken krijgt? Fraude waarbij jouw bankpin codes of creditcardgegevens worden gestolen via phishing en andere cybercriminaliteit technieken. Deze voorbeelden van 'identiteitsdiefstal' hebben een grote impact op het vertrouwen van mensen in digitale financiële diensten.

Maar het is iets dat ons allemaal vroeg of laat zal overkomen. De drempel voor het gebruik van financiële diensten wordt steeds lager, je financiële middelen worden verborgen door slechts 5 cijfers in je online banking-app of door gewoon "in de ogen te kijken" met Apple Pay. Je financiële gegevens liggen voor het grijpen als je niet van nature achterdochtig bent.

Als ik alle mails had geloofd zoals een enorme erfenis van een verre oom of whatsapp's van mijn kinderen (die ik niet heb) die hun telefoon kwijt zijn en snel wat geld willen.....

Het beschermen van onze gegevens is een topprioriteit geworden in deze wereld waarin informatie centraal staat, waar we informatie ademen. Vaak zeer persoonlijke informatie, wat een

noodzakelijk kwaad is om als persoon in onze samenleving te functioneren. Of het nu gaat om het uitvoeren van banktransacties, het online kopen van spullen op websites, het communiceren met de overheid of zelfs het sms'en naar je geheime liefde, noem maar op, (gevoelige) informatie uitwisselen vormt de basis voor de manier waarop we functioneren en ons leven leiden.

Informatie is dus kennelijk belangrijk!

Trendwatch organisaties, zoals Gartner, verwijzen naar de hoeveelheid datagroei van de afgelopen jaren. We spreken nu over hoeveelheden gegevens die alleen in volumes kunnen worden uitgedrukt, omdat het onmogelijk is deze in cijfers te lezen.

Unit	Value	Example
Kilobytes (KB)	1,000 bytes	een paragraaf van een tekstdocument
Megabytes (MB)	1,000 Kilobytes	a klein boek
Gigabytes (GB)	1,000 Megabytes	Beethoven's 5e Symfonie
Terabytes (TB)	1,000 Gigabytes	alle röntgen foto's in een groot ziekenhuis
Petabytes (PB)	1,000 Terabytes	de helft van de inhoud van alle Amerikaanse academische onderzoeksbibliotheeken
Exabytes (EB)	1,000 Petabytes	ongeveer een vijfde van de woorden die mensen ooit hebben gesproken
Zettabytes (ZB)	1,000 Exabytes	evenveel informatie als er zandkorrels zijn op alle stranden van de wereld
Yottabytes (YB)	1,000 Zettabytes	evenveel informatie als er atomen zijn in 7.000 menselijke lichamen

Onze datavoetafdruk is enorm en we vertrouwen erop dat onze IT-organisaties al deze gegevens bewaken en beschermen. Maar er bestaat geen gouden regel voor het beschermen van gegevens. Iedere organisatie probeert op haar eigen en unieke wijze te voldoen aan wet- en regelgeving voor de veiligheid en wordt helaas geconfronteerd met hun tekortkomingen op het moment dat de kaping een feit is.

Er zijn zoveel voorbeelden waarbij fantastische technologie wordt gebruikt om onze IT te beveiligen totdat een mens besluit elk veiligheidsprincipe te overtreden door een wachtwoord in te voeren dat hij of zij niet gemakkelijk kan vergeten. Het fenomeen Welcome123 heeft veel veiligheidsproblemen veroorzaakt en zal in de toekomst nog veel veiligheidsproblemen gaan veroorzaken. Eén ding is zeker: je kunt duizenden sloten op de voordeur plaatsen, maar als je de deur open laat staan, is de kans groot dat inbrekers je huis binnengaan.

17.2. DE WEAKEST LINK



Stel je dus voor dat de IT van een organisatie een kasteel is. Rondom het kasteel ligt een gracht met één toegangspunt, de hoofdpoort met een ophaalbrug. Je kunt het hek op slot doen en je haalt de ophaalbrug omhoog en je bent veilig, denk je..... Hoeveel films hebben we niet gezien of boeken gelezen waarin de vijand door de gracht kruipet en het kasteel binnengaat via het

riool of door een klein gaatje of raampje waarvan jij als kapitein van de wacht niet op de hoogte was.

Een reeks veiligheidsmaatregelen moet het kasteel beschermen:

- Gracht
- Ophaalbrug
- Muren
- Poort
- Geheime kamer
- (afgesloten) Schatkist

Maar de zwakke schakels in deze keten zijn duidelijk. Het riool of het raampje in de kasteelmuur!

Denk niet dat je veilig bent, maar hanteer het principe dat, ook al is je buitenmuur beveiligd, dit slechts een vertragingsfunctie heeft voor de vijand om je kasteel binnen te komen. Wat is de volgende hindernis die de vijand moet nemen om bij de belangrijkste schat van het kasteel te komen: jouw DATA.

De bedreiging van binnenuit, de menselijke factor: Wat als er iemand in je organisatie zit die de ophaalbrug (bewust of onbewust) laat zakken?

17.3. CLASSIFICEREN VAN DATA

Terug naar de digitale wereld van vandaag.

Om met data te kunnen werken, moet je ervoor zorgen dat je begrijpt met wat voor soorten data je te maken hebt. Welke gegevensclassificaties hebben we en wat zijn de beveiligingsregels (beveiligingsbeperkingen) die voor deze classificaties gelden?

Gegevens kunnen worden geclassificeerd op basis van het gevoeligheidsniveau (sensitivity level): **hoog, gemiddeld of laag**.

High sensitivity data (confidential)—indien gecompromitteerd of vernietigd tijdens een ongeoorloofde transactie, zou dit een catastrofale impact hebben op de organisatie of individuen. Bijvoorbeeld financiële gegevens, intellectueel eigendom, authenticatie gegevens, maar denk hierbij ook aan militaire gegevens. Om toegang tot dit soort informatie te krijgen worden aspirant medewerkers in deze organisaties gescreend.

In het geval van het kasteelverhaal zou dit het recept zijn voor het maken van buskruit en een overzicht van de buskruit inventaris en opslaglocaties.

Medium sensitivity data (internal use only)—uitsluitend bedoeld voor intern gebruik, maar als ze worden aangetast of vernietigd, zouden ze geen catastrofale gevolgen hebben voor de organisatie of individuen maar kan wel misbruikt worden. Bijvoorbeeld e-mails en documenten zonder vertrouwelijke gegevens.

In het geval van het kasteelverhaal zou dit informatie zijn over de vakantieplanning van de kasteel wachters waardoor je weet dat de bezetting van de wacht minder is.

Low sensitivity data (unclassified)—bedoeld voor openbaar gebruik. Bijvoorbeeld de inhoud van openbare websites.

In het kasteelverhaal zou dit een overzicht zijn van het kasteel, de prachtige omgeving en het aantal mensen dat in het kasteel woont.

Wie bepaalt welke classificatie van toepassing is op data? Het begint allemaal met de vraag ‘wie is de eigenaar van de data’. Met eigenaar bedoelen we de bedrijfseigenaar (als organisatie-eenheid, niet als persoon) die verantwoordelijk is voor de bedrijfsprocessen waarin deze gegevens worden gebruikt of

geproduceerd. Denk aan de verkoopafdeling waar klanten hun bestellingen plaatsen en financiële transacties uitvoeren. De records voor bestellingen in het order beheersysteem en de financiële informatie zijn doorgaans eigendom van de verkoopafdeling. De gerelateerde informatie over de inventaris en voorraad van producten is echter doorgaans eigendom van de afdeling voorraadbeheer.

Al deze informatiebronnen zijn met elkaar verbonden en afhankelijk van elkaars integriteit voor een gezonde en goed presterende bedrijfsvoering.

Dus, na identificatie van je belangrijke gegevens, moet je jezelf de vraag stellen welke maatregelen IT-organisaties moeten treffen om deze gegevens veilig te stellen, te beschermen en ervoor te zorgen dat deze beschikbaar zijn voor jouw belangrijke bedrijfsactiviteiten?

In de context van beveiliging zijn er drie belangrijke requirements voor data: Confidentiality, Integrity and Availability (CIA), in het Nederlands Beschikbaarheid, Integriteit en Vertrouwelijkheid (**BIV**):

Beschikbaarheid: De beschikbaarheid van gegevens is een maatstaf voor hoe vaak je gegevens beschikbaar zijn voor gebruik, zowel door je eigen organisatie als door een van je partners. Het is wenselijk dat je gegevens 24x7x365 beschikbaar zijn, zodat je bedrijf ononderbroken kan draaien.

Integriteit: is de nauwkeurigheid, volledigheid en consistentie van gegevens. Het wordt onderhouden door een verzameling processen, regels en standaarden die tijdens de ontwerpfasen zijn geïmplementeerd.

Vertrouwelijkheid gaat over het beperken van toegang tot informatie tot alleen de personen die deze informatie echt nodig

hebben. Vertrouwelijkheid verwijst naar de veiligheid van gegevens zoals naleving van de AVG.

Beveiliging van informatie in de overheid

In de Nederlandse Overheid wordt de BIO (Baseline Informatiebeveiliging Overheid) (verplicht) gehanteerd als standaard. Vanaf 1 januari 2019 is de Baseline Informatiebeveiliging Overheid (BIO) van kracht om de veiligheid verder te vergroten en het vervangt de baselines informatieveiligheid voor Rijk, Gemeenten, Waterschappen en Provincies.

Had voorheen iedere overheidslaag zijn eigen baseline, nu is er met gezamenlijke inspanning 1 BIO voor de gehele overheid, van BIR, BIG, BIWA en IBI naar BIO.

174. DE QUANTUM BEDREIGING

Bijna alles wat je op een computer doet, maakt gebruik van cryptografie. Dat is de reden waarom indringers meestal je e-mails niet kunnen lezen, geen toegang hebben tot je medische dossiers, geen berichten kunnen plaatsen vanaf je sociale media-account, je auto op afstand kunnen uitschakelen of met het elektriciteitsnet van je stad kunnen knoeien.

Moderne cryptografie is zo goed dat wanneer beveiligde gegevens of systemen worden geschonden, dit vrijwel nooit gebeurt doordat iemand de codering zelf heeft verbroken. De meeste inbreuken zijn het gevolg van menselijke fouten: iemand geeft per ongeluk een wachtwoord door of laat een achterdeur open in een beveiligd systeem.

Hierbij kun je denken aan moderne encryptie standaarden, zoals openbare sleutels van 2048 bits, zoals de stevigste kluizen: vrijwel onmogelijk te doorbreken, tenzij iemand een sleutel buiten laat rondslingerend. Maar het tijdperk van quantum computing zal de zaken veranderen. Een slechte techneut met een

quantumcomputer met voldoende kracht kan elke 2048-bit kluis ontgrendelen en toegang krijgen tot de gegevens die deze beschermt.

We weten niet precies wanneer quantum computers krachtig genoeg zullen zijn om 2048-bit cryptografie te kraken, maar sommige experts hebben tijdlijnen geschatst op basis van wat we tot nu toe weten. Uit het National Institute of Standards and Technology Report on Post-Quantum Cryptography blijkt dat de eerste breaches al in 2030 zouden kunnen plaatsvinden.

Neem de veelgebruikte RSA-encryptie: de publieke sleutel is een geheel getal van 2048 bits – een heel groot getal. De privé sleutel is de belangrijkste factor van dat nummer.

Maar Pinksteren en Pasen zullen op 1 dag moeten vallen voordat de krachtigste klassieke supercomputer die gebouwd gaat worden de 2048-bits sleutel zal kunnen kraken om het gecodeerde bericht te kunnen lezen.

Standaarden als RSA hebben tientallen jaren goed gewerkt omdat de mensheid simpelweg niet over de middelen beschikt om deze vormen van encryptie te doorbreken, want klassieke computers zijn beperkt. Er zijn slechts bepaalde algoritmen waarvan we weten dat ze goed werken op binaire processors. In de loop van de tijd zijn we onze samenleving gaan ontwerpen op basis van de veronderstelling dat als een probleem niet kan worden opgelost met behulp van 1-en en 0-en, het helemaal niet kan worden opgelost.

Quantum computers vertegenwoordigen een geheel nieuw rekenparadigma, waarbij binaire bits worden gereserveerd voor de complexe rekenruimtes die worden gecreëerd door het gebruik van qubits, en problemen worden opgelost die ooit onmogelijk leken.

Het is dus goed om **NU** al na te denken om je data quantum safe te encrypten. Laat je hierover eens goed informeren door experts (dat zijn wij in ieder geval niet).

De uitdaging is dat het historisch gezien lang duurt om de (cyber) security infrastructuur te upgraden en dat er daarom geen tijd te verliezen is.

NB: Alles op het gebied van data security heeft een sterke relatie met data governance. Zorg dat er goede afstemming is tussen de securityspecialisten en de data governance specialisten zodat er geen ‘gaten’ ontstaan in de beveiliging van data waardoor deze kwetsbaar kan worden.

Opdracht:

Heeft jouw bedrijf data dat nog steeds relevant is in 2030 en daarom absoluut niet in handen van derden mag vallen?

Er is zoveel te schrijven over (data)security dat het een apart boek waard is. Wij hopen dat dit hoofdstuk over security services in ieder geval de awareness bij je hebt opgewekt

Sensitive Data Discovery Services	<p>Wanneer gegevens worden opgeslagen, kunnen er gevoelige gegevens tussen zitten. Denk hierbij aan IBAN-nummers, telefoonnummers, burgerservicenummer, (e-mail)adressen etc. Sensitive Data Discovery zoekt door de hele onderneming in gestructureerde en ongestructureerde repositories naar gevoelige data. Het kan ook de gevoelige gegevens correleren. Dit is bijvoorbeeld erg handig in het geval van de AVG om alle gevoelige persoonsgegevens van één persoon te identificeren.</p>
Data Security Classification Services	<p>We hebben classificatiediensten al gedefinieerd in andere domeinen (zie Data Governance en Compliance Services). De Data Security Classification Services doen precies hetzelfde als de generieke classificatiediensten, maar dan specifiek vanuit het perspectief van databeveiliging en op een veel gedetailleerder niveau. Het gaat hand in hand met Data Discovery Services. Waar Data Discovery Services gevoelige gegevens identificeren, definiëren de Data Security Classification Services de klasse van de gegevens (CIA als voorbeeld).</p>

(Proactive) Data Access Protection Services	Deze services beschermen jouw gegevens tegen toegang door ongeautoriseerde gebruikers. Het monitort voortdurend de zoekopdrachten op de gegevens en vergelijkt deze met de profielen van gebruikers in jouw authenticatie systeem. Het kan ongebruikelijke zoekopdrachten (afwijkingen) identificeren. Op basis van beleid bepaal jij welke acties ondernomen moeten worden.
Data Security Policy Enforcement Services	Deze services zijn een toevoeging op de Data Access Protection Services door de zoekopdrachten te valideren aan de hand van beleid. Een HR-medewerker heeft bijvoorbeeld toegang tot de salarisadministratie. Dat zou geen verrassing zijn, dus de Data Access Protection Service gaat prima akkoord met dit verzoek. Maar wat als deze werknemer deze activiteit op zondag middernacht uitvoert en de salarissen van alle werknemers van het bedrijf opvraagt? Er kan bijvoorbeeld beleid worden ingesteld om het aantal opgevraagde salarissen van werknemers per dag te maximaliseren of op welke dag van de week deze opvraag kan worden uitgevoerd.

Data Security Compliance Reporting Services	<p>Financiële organisaties, overheden, zorgaanbieders, retailorganisaties, alle organisaties hebben te maken met regelgeving. De Data Security Compliance Reporting Services ondersteunen je bij het genereren van rapporten waarmee je kunt aantonen dat jouw organisatie voldoet aan de (beveiligings)regelgeving die voor jouw branche geldt.</p>
Integrated Incident Notification Services	<p>In jouw beveiligingsomgeving beschik je over een Security Information and Event Management Service (SIEM). Integrated Incident Notification Services zijn services die gerelateerde gegevensbeveiligingsincidenten verzamelen en één geconsolideerde waarschuwing naar jouw SIEM sturen.</p>
Secure Multi-party Computation Services	<p>Met Secure Multi-Party Computing (MPC)-services kun je gegevensanalyses uitvoeren op basis van verschillende bronnen, zonder de gegevens te delen. Dit is waarschijnlijk het enige deel van de diensten dat niet ergens anders gedefinieerd is, ook al zijn ze sterk gerelateerd aan (federatieve) gedistribueerde verwerkingsdiensten. Gecodeerde gegevens vereisen specifieke b.v. homomorfe encryptietechnologie om de gegevens te gebruiken zonder de noodzaak om de gegevens te ontsleutelen. Voor meer informatie over homomorfe encryptie zie: https://www.ibm.com/security/services/homomorphic-encryption.</p>

Data Encryption Services	<p>Met Data Encryption Services kunt je je gegevens coderen en je coderingssleutels beheren. Die data-encryptie kan op database niveau, op kolom-, rij- of veldniveau of op opslagniveau. Als architect moet je een reeks architectuur beslissingen over encryptie definiëren. Bijvoorbeeld 'Hou je eigen sleutel vast (HYOK)', 'Breng je eigen sleutel (BYOK)' of 'Beheer je eigen sleutel (CYOK)'. Hierbij moet je denken aan sleutelbeheer, bijvoorbeeld het gebruik van een HSM (Hardware Security Module). Zoals je kunt zien, vereist deze reeks services enige aandacht!</p>
Data Masking Services	<p>Data Masking Services zijn diensten die het maskeren van gegevens ondersteunen, zodat deze alleen kunnen worden bekeken als dit is toegestaan. Deze diensten kunnen worden opgeroepen door de Data Access Protection Services.</p>
Data Vulnerability Detection Services	<p>Data Vulnerability Detection Services detecteren allerlei interne en externe bedreigingen. Het detecteert zwakke wachtwoorden, buitensporige (DBA) logins en andere ongebruikelijke activiteiten, evenals gemiste database patches en verkeerd geconfigureerde privileges. Het detecteren van kwetsbaarheden kan leiden tot aanpassing van het huidige beveiligingsbeleid. Bijvoorbeeld dit systeem mag niet meer in een productieomgeving worden gebruikt totdat alle beveiligings patches zijn aangebracht.</p>

III. PROCESSEN EN RAAMWERKEN

*Je weet nu waar je staat, je weet wat er kan,
maar hoe kom je er?*

18. DUURZAAMHEID

Wanneer we nadenken over data-gedreven organisaties **moeten** we het onderwerp duurzaamheid wel adresseren, we kunnen het niet negeren! Zeker als je weet dat wij ook een boek hebben geschreven over de duurzaamheid van IT²³. Als je dat boek gelezen hebt dan weet je dat duurzaamheid niet alleen gaat over milieu, maar ook over het ethisch gebruik van systemen en data.



In deze editie van ons boek hebben we daarom

besloten duurzaamheid als een apart hoofdstuk te beschrijven. Niet omdat het een trend is, of omdat het beter zou ‘verkopen’ (ook al is dit boek gratis), maar omdat het echt nodig is.

Responsible Computing Framework

Wij geloven ook dat veerkracht niet kan bestaan zonder duurzaam te zijn. Als je een veerkrachtige data-gedreven organisatie wilt bouwen, moet deze duurzaam zijn. Dit om te voldoen aan de continue veranderende regelgeving op dit gebied en de eisen van het publiek.

²³ <https://github.com/OrangeSeries/Sustainability>

We hebben het Responsible Computing Framework als uitgangspunt genomen. Dit raamwerk helpt ons te concentreren op mensen, processen en technologieën binnen de context van IT. We zien dat verschillende organisaties de Sustainable Development Goals (SDG) van de Verenigde Naties als raamwerk adopteren maar wij vinden dit een te brede scope voor het IT domein.

Het Responsible Computing Framework bestaat uit zes domeinen:

1. Responsible Code
2. Responsible Infrastructure
3. Responsible Data Centers
4. Responsible Data Usage
5. Responsible Systems
6. Responsible Impact

Omdat data-gedreven organisaties IT-intensieve organisaties zijn, zijn al deze zes domeinen relevant. We bespreken hoe data-gedreven organisaties in elk van deze domeinen op een ethische en verantwoorde manier gebruik kunnen maken van schaarse middelen.

De eerste drie domeinen zijn vooral technisch van aard. De nadruk zal liggen op een geoptimaliseerd gebruik van schaarse hulpbronnen zoals elektriciteit, water en metalen. De andere drie domeinen bespreken de mogelijkheden om data-gedreven architectuur zodanig in te zetten dat deze een positieve bijdrage levert aan de samenleving.

18.1. RESPONSIBLE CODE

Om het belang van dit raamwerk voor ons onderwerp te verduidelijken, beginnen we met het domein van coderen, het schrijven van applicaties.

Tegelijkertijd is het ook het onderwerp waar de meeste winst te halen is als we kijken naar energieverbruik. AI en Big Data zijn grote energieverbruikers en veelgebruikte ontwikkel talen als Python en R zijn typische voorbeelden van zeer "dure" talen (voor wat betreft het CPU-gebruik). Python vereist bijna 80 keer meer CPU-kracht dan de oude vertrouwde taal 'C'.

Dit betekent 80! keer meer energie vergeleken met 'C' (wat een gecompileerde taal is) en 40 keer "duurder" in vergelijking met Java. Het gebruik van Python in de data en AI-wereld is onvermijdelijk, maar als je het gebruikt, gebruik het dan met beleid.

Opdracht:

Wordt Python veel gebruikt en altijd voor het juiste doel?

Worden er peer reviews gedaan op duurzaamheid?

Gebruik je duurzame ontwerp- en codeer patronen?

Maak eens een goed overzicht!

Met betrekking tot ontwerp- en codeer patronen kun je een kijkje nemen in het werk van Prof. Dr. Patricia Lago (Universiteit van Amsterdam), waar zij diverse artikelen publiceerde over software en duurzaamheid (<http://patriciaalago.nl>).

18.2. RESPONSIBLE INFRASTRUCTURE

Om code (applicaties) uit te voeren heb je IT-infrastructuur nodig. Data staat tegenwoordig op diverse plaatsen in verschillende clouds: private, public en on-premise omgevingen. Het verspreiden van je bits en bytes over het netwerk ontstaat organisch maar het bij elkaar brengen van al deze data is niet gratis, het kost energie. In een data-gedreven organisatie kunnen die hoeveelheden data groot zijn en het kost dan ook navenant veel energie om al die data telkens weer te transporteren.

Naast het transport voor de data is er ook sprake van verwerking van data. Vooral voor AI toepassingen is er behoorlijk wat rekenkracht nodig. Intelligent gebruik van CPU- en GPU-workloads kan een aanzienlijke energiebesparing opleveren. In tegenstelling tot wat je op grond van de naam zou verwachten, biedt het OpenPower-consortium een technische oplossing die niet alleen open is, maar ook gebruik maakt van een andere processor architectuur, die slechts 1/3e van de energie verbruikt vergeleken met veel van de traditionele processoren die je gebruikt in je huidige datacenter.

En tot slot: de opslag van de data, je moet het ergens kwijt. Overweeg, indien mogelijk, het gebruik van tape technologie. Een tape gebruikt geen energie als die niet wordt gebruikt en een tape cartridge bestaat voor 99% uit plastic, dat recyclebaar is. Ook compressietechnologie kan zeer nuttig zijn om de benodigde hoeveelheid opslagruimte te reduceren.

Als je deze maatregelen bij elkaar optelt kun je bij het bouwen van een nieuwe of het moderniseren van een bestaande omgeving het energieverbruik met 70 tot 90% verminderen. Of anders gezegd: je zou nog maar 10-30% van de energie die je nu verbruikt voor jouw IT-infrastructuur nodig hebben. Daar tegenover staat weer de vraag, als je de hardware levenscyclus verlengt bespaar je op grondstoffen en CO₂ uitstoot tijdens de fabricage en het transport van de producten.

18.3. RESPONSIBLE DATA CENTER

Als lezer van dit boek kunnen wij ons voorstellen dat je mogelijk geen idee hebt in welk datacenter jou data staat of waar je data verwerkt wordt. Toch zou het goed zijn om de datacenter provider te vragen naar Power Usage Effectiveness (PUE) en Water Usage Effectiveness (WUE). Of je er invloed op kunt uitoefenen of niet,

de vraag is een morele verplichting, je moet duidelijk maken dat duurzaamheid belangrijk voor je is!

Vervolgens gaan we vanuit een andere hoek naar onze data-gedreven organisatie kijken: hoe kunnen we ons project inzetten om iets positiefs bij te dragen aan onze samenleving?

18.4. RESPONSIBLE DATA USAGE

In ons data-gedreven project zorgen wij ervoor dat data veilig wordt opgeslagen, getransporteerd en verwerkt. Wij respecteren de dataprivacy van de opdrachtgever, burger of bedrijf. Wij beschrijven duidelijk hoe en waarvoor we zijn- of haar data gebruiken en doen dat uitsluitend voor het doel waarvoor toestemming is gegeven.

Wij weten precies waar welke data zich bevindt en wat het is. We beschikken over een datastrategie en een werkend governance model om de strategie uit te voeren en te bewaken.

We weten ook waar de gegevens vandaan komen en hebben de data-lineage geïmplementeerd... Toch?

18.5. RESPONSIBLE SYSTEMS

We gebruiken de gegevens in onze systemen en zijn transparant in wat we doen. We implementeren robuuste ModelOps-praktijken om "bias" en "drift" te voorkomen en we bewaren onze modellen en de gegevens zodat we te allen tijde kunnen uitleggen dat we de modellen met de juiste data hebben getraind.

18.6. RESPONSIBLE IMPACT

Wij bouwen systemen met de beste bedoelingen voor de samenleving. Met ons systeem maken wij de wereld een stukje beter. Bijvoorbeeld, bouwen wij een systeem waarin we beschikbare woningen vergelijken met een lijst van woningzoekenden. Of wij ontwikkelen een systeem dat agressie op straat signaleert en zorgen ervoor dat de politie op tijd ter plaatse arriveert om escalatie te voorkomen. Hoe kan het systeem dat jij bouwt een positieve bijdrage leveren aan de samenleving? Aan welke “Sustainability Development Goal (SDG)” levert jouw project een bijdrage?

18.7. RESPONSIBLE COMPUTING VALUES

Naast de zes genoemde domeinen zijn er ook **zes waarden** die het Responsible Computing Framework adresseert:

Sustainability: Als we naar de zes domeinen kijken, zorgen we ervoor dat duurzaamheid een mindset is bij de realisatie van al onze projecten.



Inclusiviteit: Wij bouwen een ethisch verantwoord systeem met respect voor de gebruikers.

Circulariteit: Kunnen we hergebruiken wat anderen eerder hebben ontwikkeld?

Openheid: Wij zijn eerlijk, ook als het wat minder groen is, beter eerlijk dan groenwassen! (In het Engels: greenwashing)

Authenticiteit: nemen we het echt serieus? Zit het niet alleen in ons hoofd, maar zit duurzaamheid ook in jouw hart, maakt het je echt uit? Ja, het gaat nu even om jou en niet om wat die ander doet!

Verantwoordelijkheid: Je neemt eigenaarschap, ja jij, en geen gemaar, wat is jouw rol in de data-gedreven organisatie?

Wij willen graag de nadruk leggen op één belangrijke stakeholder. En aangezien wij architecten zijn, kun je raden welke rol wij voor ogen hebben!

De (business)architect is een belangrijke stakeholder in het gesprek over duurzaamheid. Je moet duurzaamheid als niet-functionele eis meenemen in je programma van eisen, je moet minimaal één principe en één architectuurbesluit hebben dat betrekking heeft op duurzaamheid. Het is serieus: het is onze planeet en het gaat nog niet de goede kant op.

Opdracht:

Ontwerp duurzaamheid principes voor jouw IT domein. 1 principe gerelateerd aan energie en 1 principe aan ethische aspecten van een toepassing.

We begrijpen dat het niet eenvoudig is omdat er diverse (tegenstrijdige) belangen zijn. Het raamwerk ‘Responsible Computing’ heeft ons geholpen om een holistische kijk te ontwikkelen op duurzaamheid en daar zijn we erg enthousiast over.

Dus architecten:

Implementeer het principe “Sustainable by Design”.

19. RAAMWERKEN

In dit hoofdstuk adresseren wij een aantal verschillende raamwerken. Technologiebedrijven ontwerpen vaak raamwerken rondom een technologie om het succes van de inzet van die technologie te vergroten. De succesvolle inzet van een technologie hangt af van de goede afstemming van mensen (gebruikers), processen(organisatie) en de technologie. De technologie is daarbij “*slechts*” een hulpmiddel. Sommige van de raamwerken die wij bespreken zijn complementair, andere kunnen overlappen.

We geven een overzicht van het raamwerk en je zult zelf afwegingen moeten maken waar, hoe, en in welke mate je zo’n raamwerk inzet. Uiteindelijk is het raamwerk ter ondersteuning om je doel te bereiken. Een raamwerk is geen doel op zich!

19.1. BUSOPS

Dit is geen officiële term, maar een onderneming voert specifieke activiteiten uit die hen uniek maakt. Een supermarkt heeft andere activiteiten dan een kledingfabrikant of een reisorganisatie. We spreken dan veelal over de primaire processen van een organisatie. Dit domein hebben we in onze blauwdruk gepositioneerd als de bovenste laag (zie het architectuur overzicht in het hoofdstuk architectuur context). Dezebovenste laag is uniek per bedrijf en maakt gebruik van onderliggende (IT-)diensten. Door de unieke aspecten van de bedrijfsactiviteiten apart te zetten blijven er generieke diensten over. Die generieke diensten zijn veelal ook te gebruiken door de secondaire processen. Denk hierbij bijvoorbeeld aan HR, finance, procurement, etc. In de gedachten van onze blauwdruk is dit relevant omdat wij iedere dienst maar één keer willen ontwikkelen en zoveel mogelijk willen hergebruiken; een duurzame gedachte!

19.2. DEVOPS

We besteden niet veel tekst aan het onderwerp DevOps omdat er veel goed materiaal beschikbaar is, maar er zijn een paar dingen die we graag met je willen delen in de context van een data-gedreven organisatie.

Wij zijn van mening dat de DevOps aanpak de beste aanpak is voor het ontwikkelen van diensten in een veerkrachtige data-gedreven organisatie. Waarom?

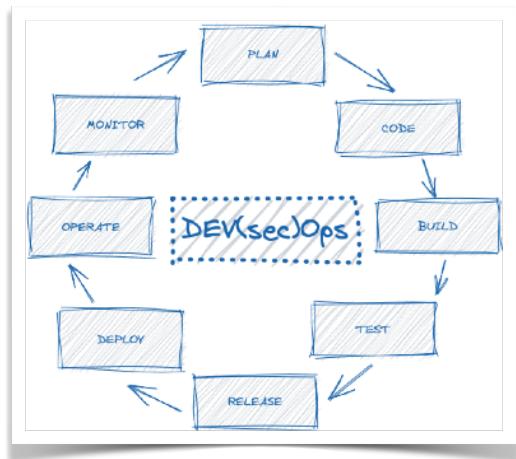
We hebben met opzet het woord veerkrachtig toegevoegd omdat je misschien denkt dat veerkracht automatisch in je organisatie wordt ingebouwd zodra je deze data-gedreven maakt. Maar is dat wel zo?

Om een voorbeeld te geven: op een douaneafdeling is er voortdurend aandacht voor het opsporen van illegaal transport van producten.

Stel dat de Douane ontdekt dat schepen van rederij De Vries, die via de haven van Singapore reizen, meer kans hebben op een illegale lading.

Deze illegale ladingen kunnen ze vanaf dat moment onderscheppen

maar na verloop van tijd zal dat gaan stoppen. Deze bedrijfsregel is dus niet langer relevant en kan afgeschaft worden, maar wat er gaat gebeuren is dat de smokkelaars alternatieve routes zoeken. Dus het vraagt continue aanpassingen van bedrijfsregels aan de



DevSecOps

situatie. In het Engels spreken we over resilience en agility, in het Nederlands hebben we daar twee mooie woorden voor: veerkrachtig en wendbaar.

Aanpassingen kunnen dus veroorzaakt worden door veranderende patronen maar er kunnen ook externe “verstoringen” optreden, bijvoorbeeld COVID of wetswijzigingen, bijvoorbeeld de EU-Act. Dit vraagt om snelle aanpassingen en flexibiliteit die alleen een veerkrachtige en wendbare organisatie kan leveren.

De applicatie die wij ontwikkelen moet natuurlijk veilig worden opgeleverd, maar één veiligheidsaspect die we explicet willen benadrukken is de kwetsbaarheid van de DevOps loop zelf (loop als in Engels).

De DevOps loop zorgt ervoor dat applicaties in verschillende stappen automatisch in productie worden gezet, het introduceert daarmee het risico dat er kwaadaardige code kan worden ingevoegd en deze “zomaar” in productie terecht komt. Daarom kom je ook DevSecOps tegen, security in DevOps inbouwen in je voortbrenging is een aan te bevelen best practice.

De loop zelf is vrij eenvoudig:

- het definiëren van het minimaal levensvatbare product (Minimal Viable Product=MVP),
- het plannen van de sprints waarbinnen code wordt ontwikkeld om (micro)services te bouwen en te testen.

Het is dus belangrijk om tijdens het testen te controleren op die kwaadaardige code. Na het testen bouw je de release en sla je die uiteindelijk op in het versiebeheersysteem, meestal het begin van een nieuwe “branch”.

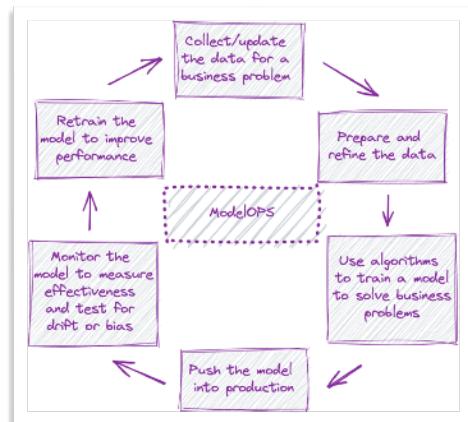
Vervolgens wordt het operationeel en vanaf dat moment is het toegankelijk voor gebruikers en kan het systeem worden gemonitord. En nu kunnen we een nieuwe cyclus starten. Hoe je

de monitoring aanpak van een uitgebreid (microservices) landschap beschrijven we in het hoofdstuk over AIOps.

19.3. MODELOPS

Zoals eerder in het boek is aangegeven: 'Fail fast, learn fast'. Dit is een fundamentele succesfactor voor een data-gedreven organisatie.

Ontwikkelaars hebben geleerd agile principes toe te passen bij het bouwen van applicaties. Dit wordt geïmplementeerd in een DevOps proces.



ModelOps

Een soortgelijk proces kun je toepassen op de ontwikkeling en gebruik van AI-modellen. Dit om de acceptatiegraad in operationele omgevingen te verbeteren en een hoger succespercentage te behalen vanuit de AI experimenten. We spreken dan van ModelOps.

Het toepassen van een beproefde aanpak of raamwerk vergroot het vertrouwen en de bruikbaarheid van de applicatie. Het bevordert de transparantie en maakt controles en validaties mogelijk. Door gebruik te maken van ModelOps kan de operationalisering steeds verder worden verbeterd, wat leidt tot succesvollere productietoepassingen.

Het werk van een data scientist is anders dan dat van een applicatieontwikkelaar. Het creëren van modellen op basis van

grote hoeveelheden data en het werken met AI en ‘onvoorspelbare’ uitkomsten is bijna het tegenovergestelde van het ontwikkelen van code voor expliciet gedefinieerde functies die kunnen worden getest aan de hand van specificaties en consistente resultaten opleveren.

De eerste vragen die gesteld moeten worden zijn: welk bedrijfsprobleem moet worden opgelost en wat moet er dientengevolge in de bedrijfsoperatie worden aangepast? Vervolgens welke applicaties moeten worden verbeterd, welke workflows worden beïnvloed? Wat kan er worden afgesproken over kwaliteitsmaatregelen voor het monitoren van de prestaties van modellen?

De praktijk leert dat de interactie tussen business en data scientists verbeterd moet worden en frequenter moet plaatsvinden om ervoor te zorgen dat de data scientist ook focus houdt op het oplossen van die bedrijfsproblemen.

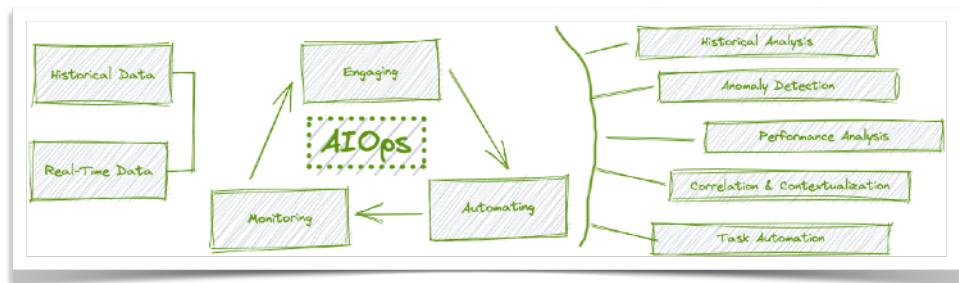
Applicatieontwikkelaars gebruiken use cases en feedbacks om te verifiëren of de requirements goed zijn geïnterpreteerd.

Data scientists kunnen de noodzakelijke data verzamelen zodra het bedrijfsprobleem duidelijk is. Dat is doorgaans niet eenvoudig want die data is veelal verspreid over de hele organisatie en soms ook daarbuiten. Daarom hebben we het architectuur domein Sorts of Data gedefinieerd. Dat is de plaats waar alle data beschikbaar is voor de data scientist om mee te werken. Zo vermijden we eindeloze discussies over de vraag of data gebruikt mag worden en voorkomen we eindeloze vertragingen bij het beschikbaar maken van de data. De data scientist kan zijn tijd besteden aan zijn kerntaak, de belangrijkste reden waarom hij of zij elke ochtend uit bed komt. Dat is het ontwikkelen en toepassen van algoritmen en modellen die inzichten bieden om de geïdentificeerde bedrijfsproblemen te helpen oplossen. Net als bij een ontwikkelen van een applicatie heeft ook het werken met modellen een iteratief karakter. Modellen worden continue

verbeterd door nieuwe inzichten toe te voegen en parameters aan te passen.

ModelOps biedt een systematische aanpak, een raamwerk, dat de end-to-end pipeline van AI-workloads ondersteunt, van training, testen tot en met de implementatie van de modellen en het monitoren daarvan. Met de juiste technologie ondersteunt het alle delen van een AI-pipeline; datasets, dataverfijning, modeldefinities, getrainde modellen, applicaties en tenslotte monitoring. Samen maakt dit lineage mogelijk en, zoals eerder vermeld, de mogelijkheid om een pipeline te reproduceren en te controleren. Op deze manier passen we het principe van fail-fast en learn-fast toe.

19.4. AIOps

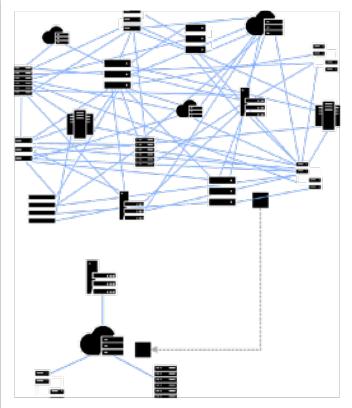


AIOps

AIOps brengt AI en Operations samen. Bij AIOps moet je in eerste instantie niet denken aan de Operations **voor** AI (AI Monitoring), maar aan de Operations met behulp van of **door** AI. Door de toegenomen complexiteit van een IT-landschap en de ontwikkeling van een microservices-architectuur exploderen het aantal services en de bijbehorende interacties. Het wordt voor mensen onmogelijk om al deze afhankelijkheden bij te houden en

de impact van een falende service op het bedrijfsproces te doorgronden. En laten we eerlijk zijn, ook in de bestaande omgeving zijn documentatie en Configuration Management Databases (CMDB's) doorgaans verouderd en geven ze niet de actuele situatie weer. Dat is de reden waarom de afgelopen jaren AI is geïnjecteerd om patronen en afwijkingen te herkennen, foutsituaties te voorspellen en oplossingen voor te stellen om incidenten (proactief) op te lossen.

AIOps verzamelt veel data door eerst een topologie van het infrastructuurlandschap op te bouwen. Je zou het kunnen vergelijken met je CMDB waar relaties tussen infrastructuur en applicaties worden opgeslagen. AIOps observeert het gedrag binnen het landschap en begint kennis op te bouwen over de relaties binnen deze omgeving. Vervolgens vergelijkt AIOps realtime gegevens met historische gegevens. De historische gegevens bevatten zowel normaal gedrag als incidenten. Wanneer er afwijkingen worden gedetecteerd, correleert het deze afwijkingen en kan dit resulteren in één of meer incidenten in je ticketing systeem. Dat is echter nog niet het einde van de cyclus, want AIOps voert ook prescriptieve analyses uit. Het kan je ondersteunen bij het oplossen van problemen en zelfs problemen voorkomen door het preventief uitvoeren van bijvoorbeeld scripts.



19.5. GARAGE

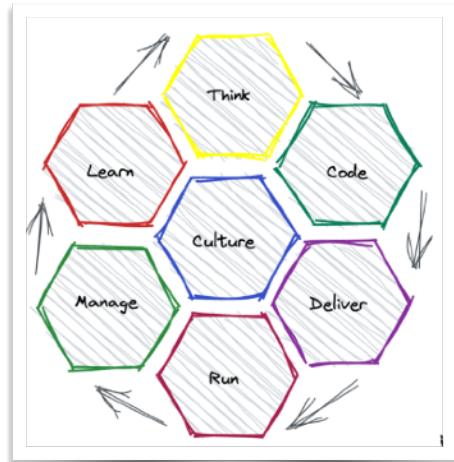
De garage is een aanpak die de verschillende processen op een pragmatische manier bij elkaar brengt. Het is gebaseerd op een multidisciplinair team van business vertegenwoordigers, ontwerpers, ontwikkelaars, data-analisten, data scientists, data-engineers, SRE's en architecten die in ongeveer 8 weken van een idee een oplossing realiseren.

"IBM Garage Is a Collection of Practices Woven Together as a Methodology To Bring all of IBM Together in Helping a Client Realize Business Value at the Fastest and Most Efficient Rate Possible."

Dr. Mohamed El-Refai.

Het startpunt is de 'Think'-fase. Op basis van de Design Thinking-methode wordt een empathymap gemaakt, worden ideeën gegenereerd en een eerste scope gedefinieerd waarbij vooral wordt gekeken naar de waarde voor het bedrijf: een zogenaamd Minimum Viable Product (MVP).

Tijdens sprints van doorgaans twee weken wordt de MVP gecreëerd en met feedback-loops worden de stakeholders geïnformeerd en kan er correctie plaatsvinden. Het MVP-product wordt geïnstalleerd in de applicatie runtime omgeving en kan in beheer worden genomen. Zodra het product in productie is, geven eindgebruikers feedback en dat kan resulteren in een nieuwe iteratie dat weer zal leiden tot een verbeterde versie van het product.



The garage method

Hoewel deze manier van werken breed wordt toegepast, is deze niet noodzakelijkerwijs goed ingebed in de gehele organisatie. Projectmanagers zijn gewend om met deliverables in tijdlijnen te werken in plaats van met sprints en backlogs van twee weken, agile versus traditionele projectmanagement milestones.

Operations teams zijn gewend om “quality gates” in hun proces in te bouwen om uitgebreide controle te houden op de kwaliteit van de productieomgeving terwijl SRE-teams hun Service Level Agreements in de loop van de tijd realiseren op een iteratieve manier. Twee totaal verschillende werkwijzen die niet makkelijk samengaan.

Vanuit het oogpunt van beveiliging en compliance moeten er maatregelen worden getroffen voordat de productie in gang wordt gezet, in plaats van een op risico gebaseerde aanpak te hanteren. Volgens een zeer bekende agile coach: “Agility moet op bestuursniveau worden begrepen en ze moeten het ermee eens zijn en erin geloven om het succesvol te maken”.

Er zijn diverse garage methoden maar om moverende reden hebben wij gebruik gemaakt van materiaal dat beschikbaar gesteld is door IBM. Meer informatie over deze Garage methode kan worden gevonden op de website²⁴ en in een artikel dat geschreven is door Distinguished Engineer Dr. Mohamed El-Refai²⁵.

²⁴ <https://www.ibm.com/garage/method/>

²⁵ <https://melrefai.medium.com/so-what-is-ibm-garage-564750a28da9>

20. DATA MESH, HET ANTWOORD OP VEERKRACHT?!

Een belangrijke vraag is: “Hoe organiseer je de beschikbaarheid van data voor jouw bedrijfsprocessen?”. Er is een interessant perspectief ontwikkeld op deze problematiek en dit wordt een **“Data Mesh”** genoemd. De term “Data Mesh” is vrij nieuw en geïntroduceerd door Zhamak Dehghani in 2019. Het idee is gebaseerd op de theorie van domain-driven design, een wat oudere theorie die momenteel weer aan populariteit wint.

Zij benadrukt het verschil tussen Systems of Record (SoR) en Systems of Insight (Sol). Aan de ene kant ziet zij dat deze type systemen steeds dichter bij elkaar komen, aan de andere kant bevestigt zij dat het nog steeds heel zinvol is om deze type systemen te blijven onderscheiden. De Data Mesh is wat ons betreft vooral van toepassing op de Systems of Insight.

Binnen een organisatie hebben we te maken met constante verandering van processen, continue repliceren van data en een diversiteit aan gebruikstoepassingen. Denk maar eens aan de opkomst van genAI; processen (moeten) worden aangepast, gebruikers gaan op een andere manier werken en er is data nodig om modellen te trainen en het ligt dan op de loer om maar even snel een copy te maken. Een Data Mesh is niet de heilige graal maar biedt wel een gedachtengoed dat kan helpen om de verspreiding van data beter onder controle te houden. Hiermee kan je organisatie regie blijven houden over datastromen en kun je de data voor jou laten werken in plaats van dat de data met jou op de loop gaat.

Traditionele data-warehouses worden gebruikt voor Business Intelligence en om rapporten te maken. Het data-warehouse is

een gecentraliseerde opslagomgeving en zodra de data de operationele grenzen van de SoR verlaat, wordt deze vaak een verantwoordelijkheid van het data-warehouse team. Sowieso een lastig verhaal, want het gaat nog steeds om diezelfde business data. Op papier was de business nog wel de eigenaar, maar die is waarschijnlijk niet altijd op de hoogte van wat er allemaal met haar data gebeurt.

Met de toenemende populariteit van AI, verschenen data-analisten en data-scientists op het podium en zij hadden andere eisen. Bijvoorbeeld ruwe operationele data in plaats van geaggregeerde en genormaliseerde data. Daarnaast hadden ze behoefte aan meer en nieuwe data om zoveel mogelijk nieuwe en betekenisvolle inzichten te verkrijgen. Omdat er data uit diverse domeinen komt is het ondoenlijk voor een data scientist om bij iedereen langs te gaan of het alstblieft mogelijk is om zijn of haar data te gebruiken. Dit betekent het volgende:

- De betekenis van de bron data moet duidelijk zijn.
- De bron data moet proactief beschikbaar worden gesteld.
- In het kader van compliance en wet- en regelgeving moet toch ook één en ander afgeschermd worden.

Dit vereist dus een beter begrip en betere beschikbaarheid van gegevens, een basis waarop vertrouwd kan worden. Wat we niet willen is dat er meerdere kopieën bestaan en we willen ook niet een eeuwigheid wachten op ons verzoek om de gegevens te krijgen die we nodig hebben voor onze analyses.

Tegelijkertijd is de complexiteit van het datalandschap toegenomen, en zelfs met data-engineers als onderdeel van een team hebben we moeite om de datadefinities bij te houden en toegang te bieden tot de data die

Het idee van een “data mesh” is om het eigenaarschap weer terug te brengen naar het bedrijfsdomein.

nodig is en gebruikt mag worden.

Een gecentraliseerde aanpak voor het beheer van data betekent, gezien het grote landschap, simpelweg dat het niet meer behapbaar is, wat leidt tot een gebrek aan goede bron- en/of domeinkennis binnen het data team.

Bovendien kan de datakwaliteit een probleem zijn. Zijn de beschikbare gegevens goed genoeg voor bijvoorbeeld de rapportage waarop de besluitvorming plaats vindt? De data voor besluitvorming moet over het algemeen van hoge kwaliteit zijn. Onder kwaliteit verstaan we onder andere dat de data accuraat, veilig en beschikbaar is. Gegevens die voor besluitvorming worden gebruikt, vereisen vrijwel hetzelfde kwaliteitsniveau als de operationele systemen. Het idee van een Data Mesh is om het eigendom van de gegevens terug te brengen naar het bedrijfsdomein om deze beheersbaar te houden. Dit wordt gerealiseerd door gebruik te maken van een domein specifiek gedreven aanpak.

De technologieën die we vanuit de Data Lake wereld kennen zijn nog steeds bruikbaar binnen grenzen van het business domein. Het domein is eigenaar van het datalandschap en heeft daarnaast de verantwoordelijkheid om afnemers uit andere domeinen (binnen de organisatie) te bedienen. Gegevens worden zoveel mogelijk beheerd in het domein waar ze worden gecreëerd en worden beschikbaar gesteld aan andere domeinen. Hoewel het in eerste instantie heel plausibel klinkt, is het nog niet zo makkelijk om dit georganiseerd te krijgen. Er kunnen twee domeinen zijn die exact dezelfde data nodig hebben, maar het eerste domein heeft het nodig voor analyse, de data mag 24 uur oud zijn, terwijl het andere domein diezelfde data real-time nodig heeft. Niet alleen de data zelf maar ook de niet-functionele eisen zijn belangrijk en voor je het weet heb je tientallen interfaces die je als

domein moet leveren en wie gaat dat allemaal onderhouden en betalen? Jij in ieder geval niet!

Dehghani heeft vier architectuurprincipes gedefinieerd:

- 1) *domain-oriented decentralised data ownership and architecture*
 - 2) *data as a product*
 - 3) *self-serve data infrastructure as a platform*
 - 4) *federated computational governance.*
-

1) Domain oriented decentralised data ownership and Architecture

Het idee is dat data-eigendom in het bedrijfsdomein ligt. Dus in plaats van een kopie van de data te verstrekken en daardoor niet te weten hoe en door wie die data wordt gebruikt, blijft het bedrijfsdomein eigenaar, ook voor het analytische deel. Dat betekent dat je binnen jouw organisatie een bedrijfsdomein hebt dat verantwoordelijk is voor alle systemen: Systems of Record, Systems of Engagement en Systems of Insight. De bedrijfskennis is geconcentreerd in dit domein.

2) Data as a product

Hoe leveren we deze gegevens aan de data scientists en data-analisten? Wij leveren data als producten en behandelen onze data scientists en data-analisten als gebruikers. Dit is te vergelijken met de Systems of Engagement waarbij we klanten en medewerkers hebben die gebruik maken van de Systems of Record. Wij leveren gegevens in het formaat, met de kwaliteitseisen, nauwkeurigheid en beschikbaarheid die ze



Data Mesh

vereisen. Wij moeten dus op de hoogte zijn van het vereiste gebruik en we houden de gebruikers op de hoogte van wijzigingen, aanvullingen etc. Om data als product te realiseren is dit een combinatie van code, data, metadata en infrastructuur.

Maar dat is best ingewikkeld!

Daarom wordt het derde principe ingevoerd:

3) Self-serve data infrastructure as a platform

Het is te veel voor elk domein om een compleet platform voor datavoorziening op te zetten, dus de beste praktijk is om één generiek platform te hebben en dat platform als een service aan de verschillende domeinen aan te bieden: dit is het *dataplatform als een service*. Het platform wordt “as a service” aangeboden door de IT organisatie. Dat wat erop draait is verantwoordelijkheid van het business domein. Hiermee zijn ook de verantwoordelijkheden voor de IT organisatie meteen duidelijk

gescheiden. Hetzelfde geldt voor de Systems of Engagement, in dat geval wordt een (Container) Platform as a Service (SaaS of PaaS) aangeboden. In het geval van de Systems of Insight zou dat een Data Platform as a Service (DPaaS) kunnen zijn. Een ‘Data Fabric’ is een mogelijke implementatie van deze service, zolang het de mogelijkheid ondersteunt om de domein georiënteerde architectuur te ondersteunen. Dit betekent dat het een multi-tenancy organisatiestructuur moet kunnen omarmen. De tenants zijn in dit geval de domeinen. Het concept van data fabric bespreken we later.

4) Federated computational governance

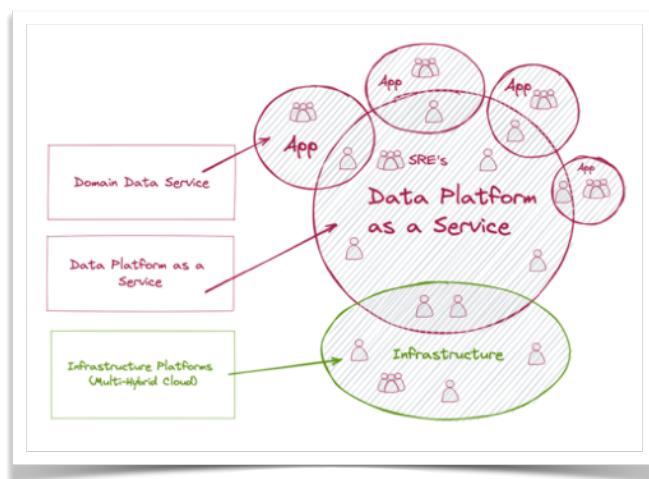
Tenslotte heb je een model nodig om samen te werken, zowel vanuit organisatorisch perspectief als vanuit infrastructuurperspectief. Vanuit een organisatieperspectief wil je weten *hoe* je welke data aan *wie* moet verstrekken, wat de prioriteiten zijn, etc. Vanuit een infrastructuurperspectief wil je het eens worden over platform standaarden, kostenallocatie, etc. Dit is het vierde principe: federated computational governance. Hier is de rol van de enterprise architect van cruciaal belang: begrijp de opties en maak de beste keuzes.

Samenvattend: Data Mesh zou wel eens de oplossing kunnen zijn om een Data Mess te voorkomen!

21. BEHEER VAN HET DATA PLATFORM MET BEHULP VAN DEVOPS EN SRE TEAMS

Nu we enkele Data Mesh principes hebben besproken, willen we de volgende stap zetten. De vraag is: "Hoe zetten we een organisatie op die een dataplatform as a service kan aanbieden en beheren?". Zelfs als je een dataplatform in de cloud gebruikt, is het nog steeds belangrijk om te begrijpen welke functies het platform biedt en welke ondersteuning er wordt geboden om die omgeving te beheren, of beter: wat er **niet** wordt geboden, want dat is wat je alsnog zelf moet doen.

In een artikel op "uncontainted.io" schetst Red Hat een organisatiemodel voor het beheer van een platform en de applicaties die daarop draaien. In dit artikel ligt de focus op een applicatieplatform, maar hetzelfde zou kunnen worden toegepast



Data platform

op een dataplatform. Laten we dit “Data Platform as a Service” noemen.

Het is georganiseerd rondom de principes van DevOps. In het platformteam zijn zowel ontwikkelaars als data-engineers, curators en infrastructuur engineers aanwezig.

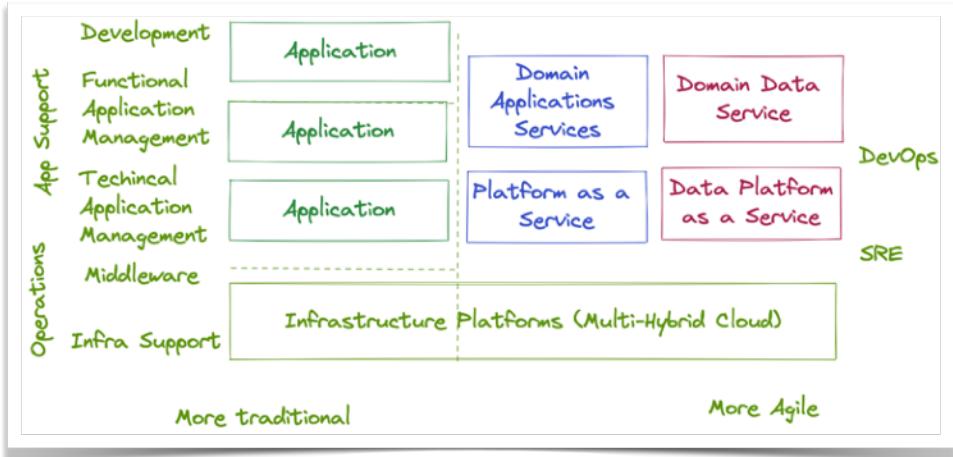
Ook al nemen deze verschillende rollen deel aan dit proces, de sleutelrol is die van de Site Reliability Engineers (SRE's). SRE's zijn verantwoordelijk voor het operationeel houden van het platform en het continu verbeteren van de platformkwaliteit. Ze zijn daarbij afhankelijk van anderen om hen te ondersteunen. Dit kunnen bijvoorbeeld de eerdergenoemde infrastructuur en data specialisten zijn.

21.1. VERSCHILLENDEN SYSTEMEN SAMENBRENGEN

In het volgende plaatje zie je op een hoog niveau hoe je de verschillende soorten systemen zou kunnen organiseren. De kolommen in het plaatje representeren respectievelijk Systems of Record, Systems of Engagement en Systems of Insight. Elk van deze systemen heeft een eigen manier van het beheer van de stack. De traditionele stack (links) bestaat uit infrastructuur beheer, veelal inclusief middleware. Daarbovenop wordt het technische beheer gedaan van de applicatie, en daar weer bovenop functioneel beheer en ten slotte de ontwikkeling. Dit is typisch de manier die toegepast wordt bij het ontwikkelen van applicaties in de Systems of Record. De nieuwe aanpak is rechts geschetst, DevOps met SRE teams voor het beheren van de hele keten.

De Systems of Insight kan gebruik maken van een dataplatform. Dit is te vergelijken met moderne, cloud native, Systems of Engagement die gebruik maken van een applicatieplatform. Je zou datzelfde model kunnen toepassen. In de praktijk zien we de behoefte om SoR op een moderne manier te organiseren en gebruik te laten maken van een cloud native applicatieplatform.

Soms wordt dit geforceerd nagestreefd en dat leidt in de praktijk nogal eens tot problemen omdat de SoR veelal zeer hoge kwaliteitseisen hebben. Denk daarbij aan een systemen dat



Verschillende soorten systemen hebben hun eigen organisatiemodel

miljoenen transacties moet verwerken of zeer hoog beschikbaar moet zijn. Ons advies is om hiermee met beleid om te gaan, het is prima dat SoR, Sol en SoE hun eigen manier van ontwikkelen en beheer hebben, als de keten maar inzichtelijk is en deze zul je als een “Business Level Objective” moeten definiëren.

22. DATA FABRIC

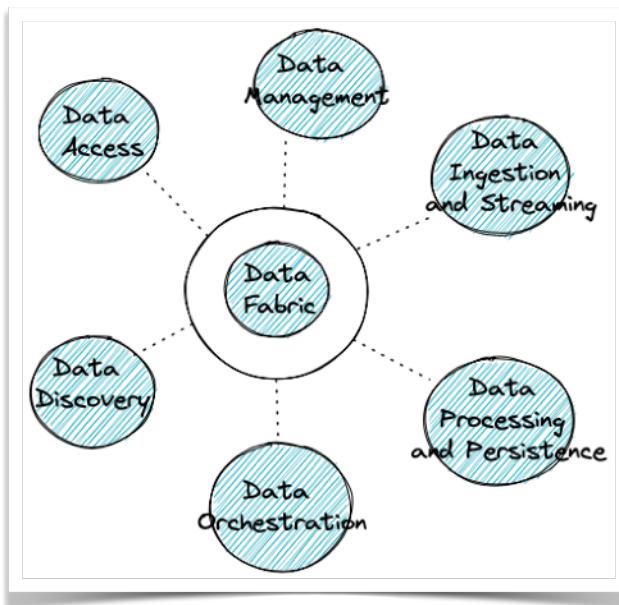
Door de jaren heen zijn er verschillende benaderingen geweest om met de steeds maar weer toenemende hoeveelheid gegevens te kunnen blijven werken. Ieder jaar verdubbeld zo ongeveer de hoeveelheid data en neemt de complexiteit toe. Combineer dat met interne organisatorische uitdagingen zoals een gebrek aan controle, een behoefte aan domeinkennis, het onvermogen om data te begrijpen en dus te bepalen welke data bij welke gebruiksscenario's passen. Het is daarmee een uitdaging geworden om de juiste informatie te verkrijgen voor besluitvorming.

In de loop der jaren hebben we verschillende verschuivingen gezien om deze uitdagingen aan te pakken. Het begon bij de ontwikkelingen van het data-warehouse concept om operationele systemen te onlasten en analyses en rapportages te kunnen uitvoeren. Je zou kunnen zeggen dat dit de geboorte was van de Sol. Vervolgens komen er oplossingen om meer data parallel te kunnen verwerken en één van deze oplossingen is een gedistribueerd model als Hadoop. We gingen ook over op het concept van een Data Lake, waarin allerlei diverse gegevens konden worden opgeslagen en georganiseerd, zodat waardevolle informatie kon worden geëxtraheerd. Je moet dat wel goed onder controle houden anders krijg je een 'data swamp'.

Technologische innovatie bood antwoorden op diverse problemen, maar desondanks blijven de uitdagingen groeien vanwege het volume en de complexiteit van data, maar ook door de hybride/multicloud-wereld. Organisaties worstelen nog steeds met het beheren en toegankelijk maken van de juiste data en gebruikers worstelen met de vraag of data überhaupt wel vertrouwd kan worden door de hoeveelheid van kopieën, problemen met de datakwaliteit etc. Research heeft aangetoond

dat de meeste data binnen organisaties niet eens wordt gebruikt²⁶ en dat de meeste data niet gedeeld wordt vanwege de zogenaamde data silos²⁷.

Een combinatie van de technische mogelijkheden die tegenwoordig beschikbaar zijn en de lessen van de afgelopen jaren uit organisatorische uitdagingen hebben geleid tot een aanpak: een **Data Fabric**. Met een Data Fabric bedoelen we niet één stukje technologie, maar een concept om bovenstaande problemen aan te pakken.



Data Fabric

26 'Rethink Data: Put More of Your Business Data to Work – From Edge to Cloud (PDF, 8.3 MB, link resides outside ibm.com), Seagate Technology, July 2020

27 "The Total Economic Impact Of IBM Garage", a commissioned study conducted by Forrester Consulting, October 2020 (link resides outside ibm.com)

Gartner definieert een Data Fabric als volgt:

“a design concept that serves as an integrated layer (fabric) of data and connecting processes. A data fabric utilises continuous analytics over existing, discoverable and inferences metadata assets to support the design, deployment and utilisation of integrated and reusable data across all environments, including hybrid and multi-cloud platforms.”²⁸

Met andere woorden, een Data Fabric kan worden omschreven als een raamwerk dat verschillende data pipelines en omgevingen met elkaar verbindt door gebruik te maken van intelligente automatisering met daarbovenop een semantische abstractielaaag.

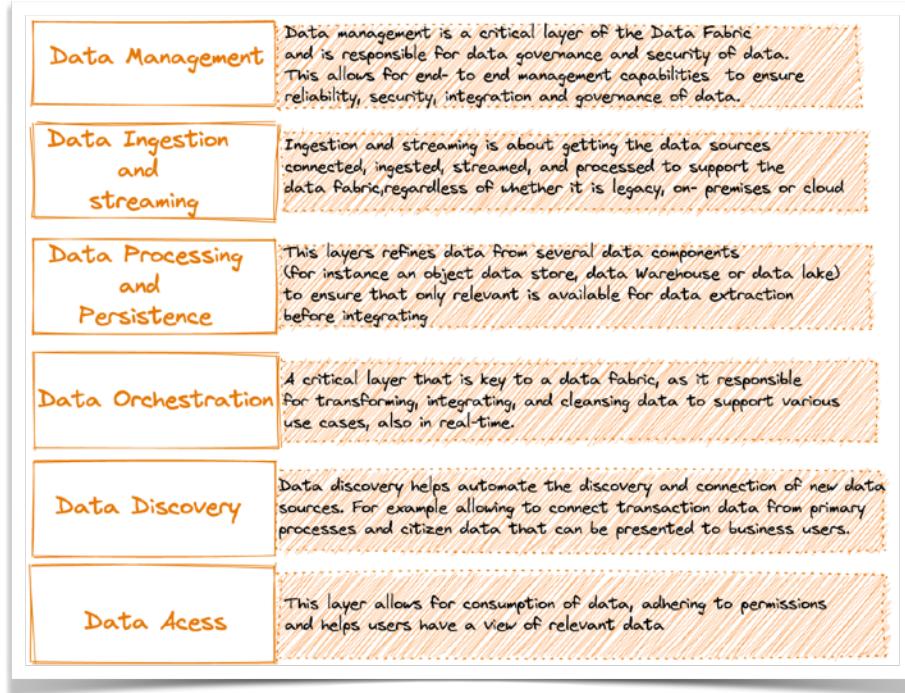
Mooie zin, al zeggen we het zelf, vooral die term semantische abstractielaaag....maar wat is dat nu eigenlijk? Op deze manier wordt de betekenis van de data duidelijk voor alle gebruikers.

Een Data Fabric koppelt de data los van de onderliggende technologie om deze aan te bieden daar waar het nodig en toegestaan is. Zo wordt data binnen een onderneming samengebracht, ongeacht de fysieke locatie van de data (of in welke silo ze verstopt zit). Met andere woorden, het verbindt de juiste gegevens, op het juiste moment, met de juiste mensen, ongeacht de locatie, door de complexiteit van de onderliggende verschillende technologieën weg te nemen.

Weer zo'n mooie zin, waar halen we het vandaan?

De semantische laag maakt het mogelijk om de verschillende silo's te verenigen, governance te verankeren en autonome zoekmogelijkheden mogelijk te maken. De sleutelfactor hier is

²⁸ <https://www.gartner.com/smarterwithgartner/data-fabric-architecture-is-key-to-modernizing-data-management-and-integration>



Data Fabric kern capabilities

metadata, en om dit te laten werken gaat het hier om technische en inhoudelijke metadata.

Deze metadata maakt tegelijkertijd privacy- en beveiligingsmaatregelen mogelijk. Het concept mag dan eenvoudig klinken, maar de uitvoering valt in de praktijk niet mee. Hoe weet je nu als gebruiker van de (meta)data wat de exacte betekenis is van een begrip en waar het voor gebruikt wordt? Een goede definitie van gegevens en zoekmogelijkheden vanuit een zelfbedieningsperspectief geven de mogelijkheid om de betekenis van data beter te begrijpen. Hierdoor kunnen bijvoorbeeld diensten aan burgers gerichter worden aangeboden of kan het concurrentievermogen worden vergroot door betere inzichten te genereren die weer kunnen leiden tot betere

beslissingen.

Naast deze semantische laag is data virtualisatie een andere belangrijke capability van een Data Fabric. Het maakt het mogelijk om verbinding te maken met verschillende bronnen en deze samen te brengen zonder een traditionele kopie of verplaatsing. Dit realiseren we door een virtuele laag te creëren die gebruikt kan worden door analisten of data scientists. Het betekent niet dat het ETL (Extraction, Transformation and Load) vervangt; elke technologie heeft zijn plaats afhankelijk van de use case en/of niet-functionele vereisten.

Forrester geeft een mooie opsomming van de key capabilities die noodzakelijke zijn voor architectuur en implementatie²⁹:

Samenvattend: Data Fabric biedt intelligente automatisering, laat data op z'n plek, brengt ML/AI naar de data en biedt toegang tot de juiste data.

Wat betekent dit in relatie tot een Data Mesh, dat in een eerder hoofdstuk is besproken? In onze optiek heeft elk zijn eigen plaats, een Data Fabric concept zou bijvoorbeeld kunnen worden toegepast in een Data Mesh of vice versa.

Bijvoorbeeld, een Data Fabric biedt een semantische laag die ingezet kan worden die om een federated computational governance te implementeren. De data in een Data Mesh omgeving kan weer als brondata in een Data Fabric worden gebruikt.

²⁹ Forrester: "Enterprise Data Fabric Enables DataOps"

23. VERTROUW JIJ OP GENAI OM BESLISSINGEN VOOR JOU TE NEMEN?

Generatieve AI bestaat al een tijdje, maar kreeg onlangs met de introductie van ChatGPT veel aandacht. ChatGPT is een geweldige tool dat vragen kan beantwoorden of taken voor je kan uitvoeren en is gebaseerd op AI-technologie. De vraag voor ons is of en hoe we die technologie kunnen inzetten om onze organisatie nog meer data-gedreven te maken. Moeten we nieuwe diensten toevoegen of valt het onder de bestaande AI-mogelijkheden die we eerder hebben besproken, of speelt het helemaal geen rol in een data-gedreven organisatie en valt het daarmee buiten onze scope?

Laten we eerst in eenvoudige woorden uitleggen hoe ChatGPT in de basis werkt en waarvoor het kan worden gebruikt.

ChatGPT maakt content en wordt daarom generatieve AI genoemd. Het is AI die iets nieuws creëert! Maar is dat wel echt nieuw? Een LLM is een model dat is getraind met specifieke informatie. Dit zijn grote hoeveelheden! Op basis van de dataset waarmee het model wordt getraind kan het ingezet worden voor specifieke doeleinden. In het geval van ChatGPT is het getraind met algemene internet informatie waardoor je dit model voor generieke doeleinden kan gebruiken. GenAI is gebaseerd op een Large Language Model (LLM). Een LLM is een vrij eenvoudig concept. Het voorspelt uitsluitend en alleen het meest waarschijnlijke volgende woord in de zin. Het kijkt naar jouw vraag en voorspelt het meest waarschijnlijke woord dat zou moeten volgen. Dat is simpel! Het concept is heel eenvoudig, maar die berekening is gebaseerd op enorme hoeveelheden

gegevens en daarom buitengewoon krachtig. Er zit een enorme hoeveelheid ‘kennis’ in die data.

Tegelijkertijd is dit de beperking. Het beschouwt de data waar het mee getraind is als de waarheid, en over het algemeen is dit juist, vooral als het feitelijk is. Maar als je vragen stelt over politiek, religie, cultuur, of meningen dan kunnen de antwoorden vanuit jouw persoonlijke perspectief twijfelachtig zijn. Een Large Language Model heeft namelijk altijd een antwoord. Dit kan leiden tot iets dat we hallucinatie noemen, het uitkramen van onzinnigheden. Maar die onzinnigheden brengt hij wel op een overtuigende wijze!

Het model gaat over taal, dat kan een natuurlijke taal of een programmeertaal zijn.

We hebben wat tests gedaan en laten we de resultaten bespreken.

23.1. LLM VOORBEELDEN

De eerste vraag die wij een LLM stelde was om een inleiding voor ons boek te schrijven:

Can you generate as short introduction for a book called blueprint for data-driven organisations

In an increasingly data-centric world, organizations are discovering that the true power of success lies in their ability to harness and leverage the vast sea of information at their disposal. Welcome to "Blueprint for Data-Driven Organizations," where we embark on a transformative journey that uncovers the keys to building a thriving and agile enterprise through data.

This book is a comprehensive guide that reveals the essential principles, strategies, and practical insights to unlock the potential of data and elevate your organization to new heights. From understanding the value of data as a strategic asset to implementing cutting-edge analytics, we delve into the core principles that pave the way for data-driven decision-making, efficiency, and innovation.

Join us as we navigate through real-world case studies, best practices, and expert advice from industry pioneers, all aimed at equipping you with the tools needed to foster a data-enabled culture within your organization. Whether you are a seasoned data professional or a curious business leader, this blueprint is designed to empower you to turn raw data into actionable intelligence and steer your organization towards unparalleled success in today's dynamic landscape.

Let's embark on this data-driven journey together, as we redefine what it means to thrive in an era where data is king. Get ready to unlock the full potential of your organization and chart a path to sustained growth, efficiency, and innovation. Your blueprint awaits!

Geweldig, vind je niet? Wij zouden het niet beter kunnen. Het mist misschien onze persoonlijk kwinkslag en is een beetje te "perfect". Maar desalniettemin, inhoudelijk is het juist. Laten we een ander voorbeeld proberen

Can you generate a very funny short introduction for a book called blueprint for data-driven organisations?

Het resultaat was een stukje tekst en het zag er op het eerste gezicht goed uit. Toen we het echter beoordeelden, was het niet waar we om vroegen, omdat de gegenereerde tekst uitging van een grappig boek. We vroegen om een grappige introductie, maar in de tekst stond dat het boek grappig was. Ook al hebben we een toegankelijk boek geschreven, het is niet bedoeld als grappig. Wij willen gewoon een grappige inleiding tot het boek. Dit is een goede weergave van hoe het LLM werkt, deze nuances in context zijn moeilijk.

Vervolgens hebben we nog twee vragen gesteld:

What is the best example of a data-driven organisation?

What is a data-driven organisation?

De gegeven antwoorden waren zeer nuttig. Het antwoord op de tweede vraag was een lijst met 10 punten waarop een data-gedreven organisatie zich zou moeten richten.

En dit was de samenvatting:

Over het geheel genomen maakt een data-gedreven organisatie gebruik van de kracht van data om een concurrentievoordeel te behalen, processen te optimaliseren en betere producten of diensten te leveren, waardoor de algehele prestaties worden verbeterd en de strategische doelstellingen worden bereikt.

Een krachtige samenvatting, nietwaar? Je kunt er zeker over discussiëren, maar het lijkt een goede definitie. Maar waar komt het vandaan? Niemand weet het, tot een auteur “zijn” tekst terugleest. Wie is nu de eigenaar van deze definitie? Wij, de auteurs van dit boek?

23.2. LARGE LANGUAGE MODELS IN EEN DATA-GEDREVEN CONTEXT

Ook vroegen we ChatGPT hoe het ingezet kan worden in data-gedreven organisaties. Het gegeven antwoord was echter een antwoord op de vraag hoe AI en niet GenAI toegepast zou kunnen worden in data-gedreven organisaties.

ChatGPT kan een goed hulpmiddel zijn om als chatbot te gebruiken, het is niet echt bruikbaar als input voor je bedrijfsbeslissingen. Het kan wel menselijke beslissingen ondersteunen.

Ons standpunt is om GenAI te gebruiken in je processen ter ondersteuning, maar niet om je processen direct te sturen.

Zie het hoofdstuk Data Analysis Services voor verantwoord gebruik van AI.

24. DATA-DRIVEN PROGRAMMING

Waarom hebben we dit onderwerp in ons boek opgenomen? De context van data-gedreven programmeren verschilt van de context van data-gedreven organisaties, maar het paradigma is vergelijkbaar. In een inspirerende bijeenkomst met *mainman/core-origens* bespraken we de toepassing van data-gedreven programmeren voor de Systems of Record. Naast het paradigma heeft data-gedreven programmeren ook te maken met IT en ervan uitgaande dat de lezers van ons boek voornamelijk IT-mensen zijn, zou deze vergelijking kunnen helpen om data-gedreven organisatie beter te begrijpen. Het heeft ons een interessant inzicht gegeven dat we graag met je willen delen in dit hoofdstuk.

Data-gedreven programmeren is het tegenovergestelde van commando gestuurd programmeren. Commando gestuurd programmeren is procedureel. Het voert een opdracht uit, die vervolgens een volgende opdracht uitvoert, enz.

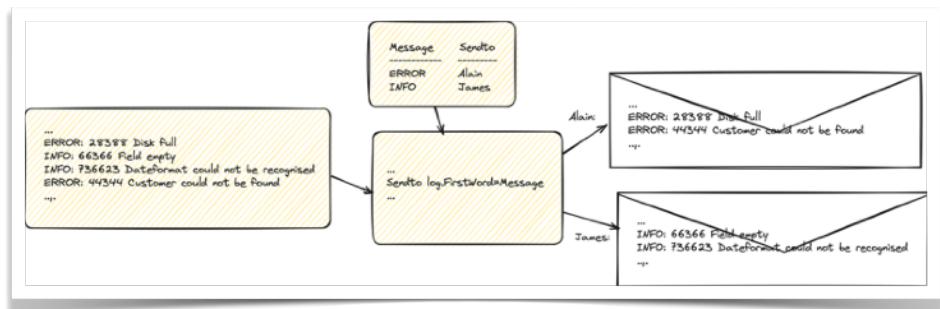
Data-gedreven programmeren analyseert de data en voert logica uit op basis van de informatie die wordt meegegeven. Dit klinkt logisch, maar laten we een eenvoudig voorbeeld gebruiken om het duidelijk te maken.

In ons voorbeeld gebruiken we de verwerking van een logbestand. Een logbestand bevat berichten over het functioneren van een applicatie. Dit kunnen informatieve berichten zijn, maar ook foutberichten. Die informatieve informatie kun je sturen naar een programmeur, hij moet het morgen oplossen, maar de informatie over de fouten gaan naar

operations want het probleem moet onmiddellijk worden opgelost.

Je kunt een programma schrijven dat de ‘Error’ verstuurd naar de operations persoon en de info naar de programmeur.

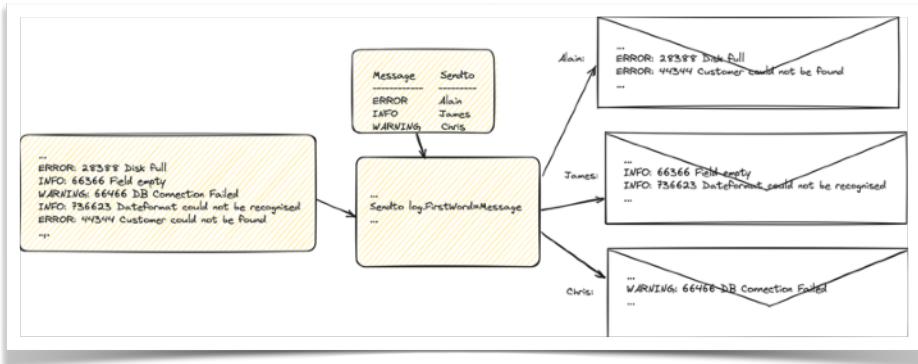
Op de afbeelding is te zien dat wij op basis van de gegevens berichten naar verantwoordelijken sturen.



De data stuurt de logica bij data-gedreven programmeren

Data-gedreven programmeren is flexibel. Je hoeft je programma niet te wijzigen als de gegevens veranderen. Als er bijvoorbeeld een nieuwe foutcategorie, in dit geval “Waarschuwing”, wordt toegevoegd, hoef je alleen je input data bij te werken!

Tegelijkertijd wordt de leesbaarheid van je programma wel wat minder omdat in het eerste geval je in het programma zelf kan zien welke foutboodschappen er waren namelijk INFO en ERROR. In het data-gedreven programmeren moet je zowel naar het programma als naar de data kijken om te begrijpen hoe het precies werkt. Nu is dit een eenvoudig voorbeeld maar ik heb meegemaakt dat iemand een programma maakte dat zo flexibel was dat de input een eigen programmeertaal werd. Dat is niet de bedoeling en de kunst is de balans te bewaren.



XML en JSON files zijn andere voorbeelden van data-gedreven programmeren. Ze kunnen heel veel informatie bevatten op basis waarvan de applicatie beslissingen neemt. De XML of JSON files kunnen zo complex worden dat er weer een programma geschreven wordt om deze files te genereren!

Tot zover data-gedreven programmeren, hopelijk is de essentie hiervan duidelijk. Laten we dit paradigma nu eens toepassen op onze data-gedreven organisatie. In plaats van de applicaties uit het voorbeeld nemen wij de bedrijfsprocessen als uitgangspunt. Wat voor bedrijfsprocessen hebben jij in je organisatie? Zijn het standaard processen die altijd dezelfde taken uitvoeren of zijn het processen die worden aangestuurd door data? In het eerste geval is het proces voorspelbaar en eenvoudig, maar heb je veel processen, ieder proces doet zijn eigen ding. In het tweede geval zijn je processen flexibeler, heb je er minder en laat je ze sturen door je data. Dat is het idee van een data-gedreven organisatie!

Business analysten hebben de belangrijke verantwoordelijkheid om de optimale balans te vinden tussen flexibiliteit en complexiteit van het modelleren van bedrijfsprocessen.

Data-gedreven organisaties leveren niet alleen de data, ze moeten ook data-gedreven processen implementeren om met data om te gaan. Je moet definiëren welke data je nodig hebt om je processen te sturen, welke logica je onderbrengt in processen, welke logica je onderbrengt in bedrijfsregels en welke logica je hard vastlegt in je bedrijfsproces. Om de juiste keuzes te maken zijn twee dingen belangrijk, ervaring en kennis van de werking van je organisatie. En niet alleen hoe je organisatie vandaag de dag werkt, maar ook de gewenste toekomstige werkwijze.

Data-gedreven organisaties moeten niet alleen data beschikbaar stellen maar ook hun processen zo inrichten dat ze het maximale uit deze data kan halen

Van proces gestuurd naar data gestuurd

Om een en ander nog eens naast elkaar te zetten hebben we de uiterste opgenomen in de onderstaande tabel. Het gaat erom dat je het onderscheid kan maken, in de praktijk kom je de extremen niet tegen.

Proces-gedreven	Data-gedreven
Proces is uitgangspunt: data de variabele	Data is uitgangspunt: proces volgt de data
Eén proces per doel.	Eén proces dient meerdere doelen.
Als het handig is kopieren we een gedeelte van de data voor een ander proces. Data volgt.	Data kwaliteit is essentieel. Master data als de single source of truth.
Doe de dingen op de juiste <u>manier</u>	Doe de juiste dingen op basis van variabele data
Bescherm je applicatie/proces!	Bescherm je data!

Is dit een paradigma verandering voor jou of voor jouw organisatie? Dan komt er mogelijk meer bij kijken dan je in eerste instantie dacht!

25. EN VERDER

Na al deze onderwerpen waar je als weerbare en wendbare data-gedreven organisatie in meer of mindere mate mee te maken hebt of krijgt is de vraag: hoe verder? Het is geen boek waarin je begint met stap 1 en eindigt met stap 10 en voilà, je bent een data-gedreven organisatie. Er zijn veel keuzes te maken, veelal zijn ze complementair, soms staan ze tegenover elkaar. Misschien vind je het allemaal onzinnig en heb je niets aan dit boek. Afhankelijk van je emotionele status besluit je het boek weg te gooien, de open haard er mee aan te maken of je geeft het aan iemand anders en ga je op zoek naar je eigen wegen om je organisatie data-gedreven te maken. Dan hebben we in ieder geval één ding bereikt, je hebt zelf een mening gevormd!

Wanneer we echter iets hebben geadresseerd dat je aanspreekt is de vraag: hoe verder? De hoeveelheid capabilities overondert je misschien. Daar kunnen we inkomen en onze aanpak is dan ook niet om op de eerste dag een omgeving op te zetten die alle capabilities implementeert. We voorzien een geleidelijke aanpak.

De beste start is om een business scenario te selecteren, liefst één die niet al te complex is als eerste, maar wel toegevoegde waarde heeft voor het bedrijf. De tweede stap is dan om de bijbehorende use case te ontwikkelen met behulp van de services die je nodig hebt om deze te implementeren en te operationaliseren. Dit is ook het moment dat je kunt overwegen om de eerdergenoemde “garage methode” te gaan inzetten om een MVP te bouwen. Waarschijnlijk is het ook het moment dat

problemen gaan beginnen en je allerlei vragen hebt of krijgt waar je een antwoord op moet zoeken.

Perfect! Die vragen zijn belangrijk. Je moet ze analyseren en zodanig beantwoorden dat je de oplossingen ook voor toekomstige projecten kunt inzetten en ze tot standaarden kan promoveren.

Dit is het moment waarop je de A0(841 x 1189) blauwdruk poster kan gaan gebruiken. Je kunt op de plaat aangeven welke services je nodig hebt. En als je poster volgeklad is kun je een nieuwe downloaden vanuit [github.com](https://github.com/datadrivenblueprint/downloads#readme):

<https://github.com/datadrivenblueprint/downloads#readme>

Je vindt op die lokatie zowel de laatste versie van dit boek als de poster.

Je zal thuis geen A0 printer hebben, maar wellicht kan de repro-afdeling van jouw organisatie hierbij helpen of kun je van een externe print-dienst gebruik maken.

Mogelijke vragen waar je mee te maken krijgt:

1. Ik heb een service geselecteerd en nu moet ik de keuze maken voor een implementatie. Heb ik deze service al en kan ik deze ook gebruiken in de context van mijn use case, of moet ik misschien een nieuwe service ontwikkelen die ook rekening houdt met toekomstige behoeftes?
2. Ik moet een methode ontwikkelen, bijvoorbeeld ModelOps, maar heb deze capabilities niet. Welke capabilities heb ik precies nodig en kan ik klein beginnen? Welke adviseurs kunnen mij hierbij helpen?

3. Ik heb data nodig waar ik geen eigenaar van ben, hoe kom ik aan die data?
4. Ik heb een interface nodig om bestanden over te halen. Het zou gewenst zijn om een file transfer service te hebben, maar mijn organisatie heeft alleen een API integratie mogelijkheid. Ga ik deze gebruiken of ga ik toch voor een optimale oplossing (bijvoorbeeld een Data Fabric concept) omdat ik op termijn meer toepassingen hiervoor zie?
5. Hoe integreer ik de bestaande situatie met de toekomstige, rekening houdend met people, proces en technology.

25.1. AAN DE SLAG

Je kunt nu de poster pakken en de use cases op de plaat plotten. Geef de services die je gebruikt in je use case een uniek nummer. Je kunt dit voor meerdere use cases doen. Bijvoorbeeld de use cases binnen je EPIC. Na afloop van deze werkzaamheden heb je een compleet overzicht van de benodigde services. Denk erom dat *de functies* die geleverd worden door één service wel per use case kunnen verschillen.

Vergeet ook niet om rekening te houden met de onderliggende infrastructuur services om je use cases “enterprise ready” te maken zoals we dat noemen. Daarnaast heb je te maken met een hoeveelheid van stakeholders die variëren van business, strategie, architectuur, operations en diverse technische experts. Zorg ervoor dat de stakeholders zich herkennen in jouw oplossing!

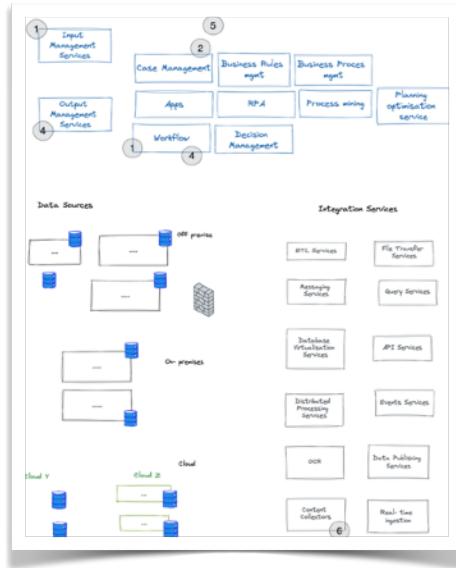
25.2. NOTITIES

Use case 1

Use case 2

Hieronder vind je een voorbeeld hoe je de use cases kan weergeven met nummers. We hebben gemerkt dat het heel rommelig kan worden in de praktijk om je use case op de volle plaat weer te geven. In eerste instantie is dat wel heel handig omdat de services als een checklijst fungeren, maar om er op een overzichtelijke wijze mee verder te werken kun je het beste de services die je niet gebruikt achterwege laten.

Nu je het kun je gaan over de je services. kun je niet puntjes hier en daar pragmatische moeten. Je wacht tot compleet is kun je



overzicht hebt nadenken realisatie van Ongetwijfeld alles tot in de uitwerken en zul je oplossingen bouwen. Als alles uitgewerkt wachten tot je

Use case voorbeeld

een ons weegt.

Think big, act small is hier van toepassing. Er zijn complete platformen beschikbaar (in de cloud) die je kunt gebruiken om te versnellen.

Als je een datacenter manager of een architect en je overweegt om een dataplatform te ontwikkelen, zou je eens over het volgende kunnen nadenken.

Als je bovenop een data platform dan ook nog business automation, een applicatie ontwikkel platform en een integration platform kan leveren dan heb je een set van diensten geleverd om een wendbare en flexibele data-gedreven organisatie te realiseren. Deze drie platformen zijn beschikbaar in de markt en als je dit aan kunt bieden kunnen de ontwikkelaars zich richten op het realiseren van business requirements.

EPILOOG

Er valt veel te zeggen over het bouwen van een data-gedreven organisatie, zoveel dat je hele enterprise content management systeem kunt vullen met honderden boeken en duizenden artikelen. De toepassing van AI in een data-gedreven organisatie is onvermijdelijk. We zien dat veel bedrijven trends in de richting van cloud adoptie en/of microservices-architectuur volgen op basis van hun strategie voor “Systems of Engagement” zonder zich de impact voor de “Systems of Insight” te realiseren. Je kunt ze niet los van elkaar zien maar ze *zijn niet hetzelfde!* Ze hebben andere karakteristieken, andere eisen, andere patronen, hou daar rekening mee.

Een data-gedreven organisatie moet in ons perspectief eigen principes definiëren, zelfstandig beslissingen kunnen nemen en eigen processen ontwikkelen zoals MLOps. Vragen die je jezelf moet stellen en die onderdeel zijn van typische Sol problematiek:

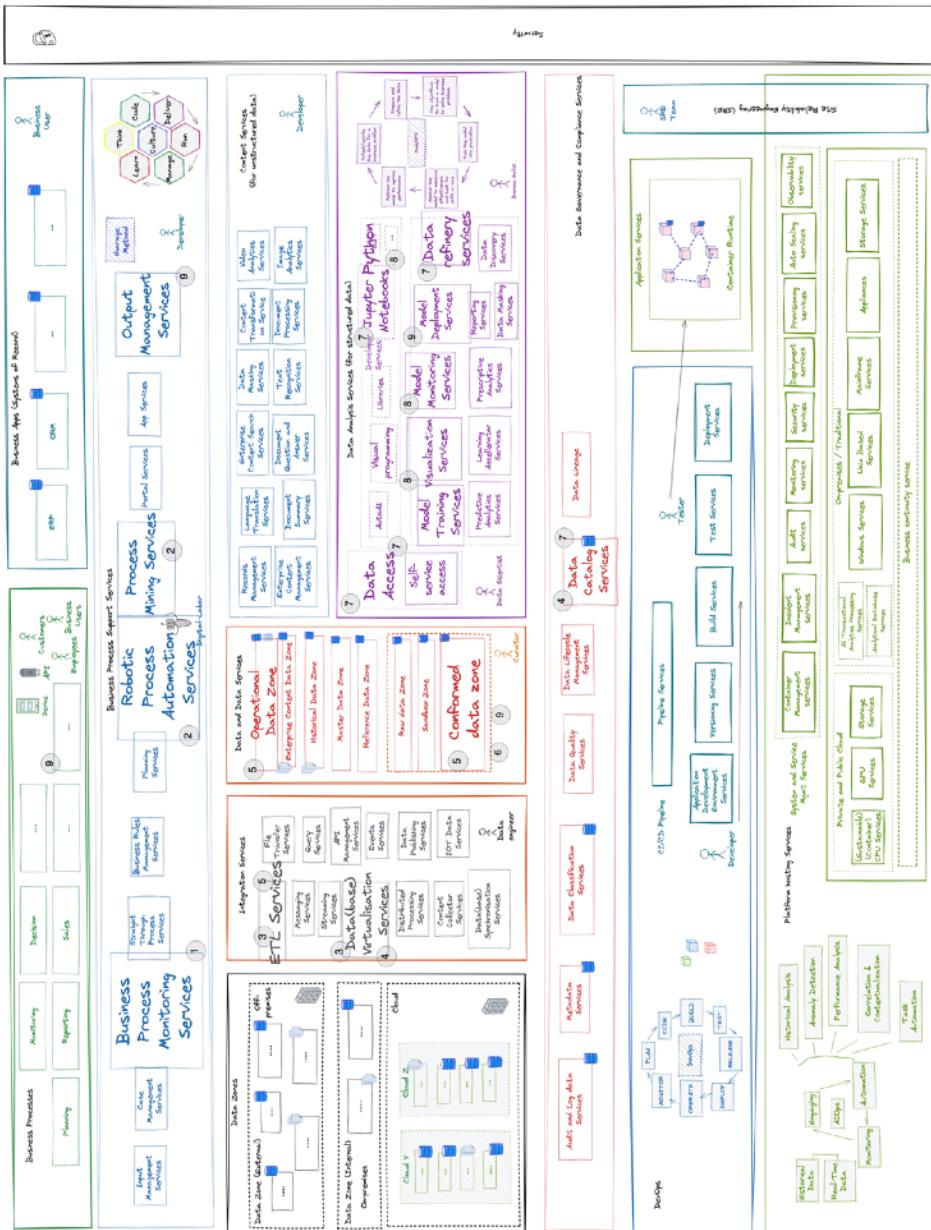
- Wat zijn de kosten van het plaatsen van grote hoeveelheden gegevens in de publieke cloud? Wat zijn de kosten voor het gebruiken of migreren van de gegevens uit de cloud (egress)?
- Is agile de manier van werken voor betrouwbare AI?
- Hoe verklaar je “duurzaamheid” als 25% van je servers analyses uitvoert?
- Wie neemt de beslissingen? Je applicatieontwikkelaars, je data scientists of je governance-organisatie?
- Site Reliability Engineers, bestaan ze alleen in naam of ook in de praktijk?
- Vertrouw je de data? **Of** je vertrouwt de data!
- Is data-gedreven een keuze of een verplichting?
- Wat als ik analyses op mijn operationele gegevens zou kunnen uitvoeren?
- Willen we transparant zijn of is dat een moetje?

- Tot hoever willen we afhankelijk zijn van onze “Systems of Insight”? Wat betekent dat voor de kwalitatieve aspecten? Zijn wij bereid de prijs te betalen?

Er is niet een juist antwoord, maar dit zijn stuk voor stuk goede vragen om te stellen! Inmiddels, aan het eind van dit boek, hebben we elkaar een beetje leren kennen, we weten dat jij ook andere (en betere) vragen hebt! Dat is mooi, want daarmee hebben we ons doel bereikt: jou een holistisch overzicht geven van de impact van het bouwen van een data-gedreven organisatie, *nadenken* over de impact van technologie en bespreken welke services er nodig zijn.

Wij hebben ervan genoten, bedankt voor je interesse!

Laila, Jan, Ronald



APPENDIX

En nu aan de slag met het materiaal. Het is eenvoudig om een Quick start te maken. Hou het vooral simpel in het begin want je wil niet meteen de oceaan proberen te koken.

Wij hebben een drie-stappen plannetje waar je mee kan beginnen:

- Beschrijf de use case in stappen
- Map deze stappen op de services
- Leg de relaties tussen de services

In dit geval hebben we de relatie met de infrastructuur achterwege gelaten om het voorbeeld eenvoudig te houden. Afhankelijk van jouw doelstelling kun je dat ook doen, maar als je met je operations afdeling praat is het juist verstandig om de infrastructurele en operationele services er wel bij te nemen. Uiteindelijk moeten zij weten op welke infrastructuur het moet gaan draaien en hoe ze dit moeten gaan monitoren en beheren.

- 1 Process owner Marie receives a notification that the waiting times for requests are increasing which needs her attention.
- 2 She runs a simulation, based on the expected numbers ,and finds out that the number will only increase prompting her to take immediate action as her department will not be able to handle the number of requested within the required timelines. She reaches out to the privacy officer to request approval for the usage of a machine learning model of machine learning to implement a scoring function which will separate out special cases whereas regular non- risky cases can be processed directly. David, a data scientist is assigned to work on the project.
- 3 David, the data scientist cannot wait to start! He wants to build a solution that separates the low risk request from request that require personal attention. These low risk request can then be handled automatically!
- 4 He goes to the catalog and self-service portal to request the required data.
- 5 Meanwhile Marco, the data engineer receives a request to provide the data for David in the operational environment.
- 6 Marco makes the data available in the sandbox zone, masking private data and accessible for David and Marie.
- 7 David prepares and transforms the data before applying Machine Learning.
- 8 After a couple of iterations the model seems to return the proper results and he shares the outcome with Marie. They review the model as well as the outcome.
- 9 The model will be approved and operations will implement the model for production.

GLOSSARY

A0	Het basis papierformaat	Het A-Formaat is een papier formaat dat gebruikt wordt in landen waar het metrische systeem van toepassing is. A0 is het basisformaat van exact 1m2. Als je het papier 5x zou vouwen krijg je het A5 formaat, dat is het formaat van dit boek!
AI	Artificial Intelligence	Computerintelligentie die verder gaat dan de standaard IF THEN ELSE-bewerkingen van een computer.
AIX	Advanced Interactive eXecutive	Serie Unix-besturingssystemen, ontwikkeld door IBM. Ze worden meestal gebruikt voor bedrijfsservers.
AKS	Azure Kubernetes Services	Een door Microsoft ontwikkelde en beheerde Kubernetes implementatie.
API	A Programming Interface	Een manier waarop applicaties met elkaar kunnen communiceren en waarmee bedrijven zich kunnen openstellen voor anderen.
CDN	Content Delivery Network	Een netwerk van servers die een nomadisch leven leiden en tegelijkertijd samenwerken om hoge beschikbaarheid en prestaties te bieden.
CDO	Chief Data Officer	De persoon die verantwoordelijk is voor alles wat met data te maken heeft. Een relatief nieuwe rol in organisaties, populairder geworden met de introductie van GDPR. Als je geen CDO hebt stel er dan een aan.
CI/CD	Continuous Integration / Continuous Delivery	Een werkwijze om continue applicaties te integreren en implementeren.
CIA	Confidentiality, Integrity and Availability	Een model dat richting geeft aan het beleid voor informatiebeveiliging binnen een organisatie, waar we geen grappen over willen/mogen maken.
CIO	Chief Information Officer	De baas van de IT afdeling, waarschijnlijk ook die van jou!
CISC	Complex Instruction Set Computer	Minimaliseert het aantal instructies per programma. Het tegenovergestelde van RISC.
CMDB	Configuration Management Database	Een database om de hardware en software en hun relaties bij te houden.
COTS	Commercial off the Shelf	Een product dat je bij een leverancier koopt en aanpast om het bruikbaar te maken in jouw eigen organisatie.
CPU	Central Processing Unit	Het brein van de computer.

CSMO	Cloud Service Management and Operations	Niet de populaire Chief Social Media Officer, maar een raamwerk voor het draaien van applicaties in de cloud.
CSP	Content Services Platform	Evolutie van document mgt naar ECM naar CSP.
CTO	Chief Technology Officer	De baas van de technologiestrategie, inclusief IT.
DBA	Database Administrator	Degene die ervoor zorgt dat databases worden ontworpen, onderhouden en uitgevoerd zoals vereist. Oh, en hij/zij zorgt ervoor dat iemand ook toegang heeft tot de data(base)...
ECM	Enterprise Content Management	Proces voor het beheren van de levenscyclus van content (van documenten tot video) en de bijbehorende technologie.
ERP	Enterprise Resource Planning	Software die wordt gebruikt voor dagelijkse handelingen, zoals transacties, inkoop of boekhouding.
ETL	Extract, Transform, Load	Een manier om gegevens uit meerdere bronnen te combineren en beschikbaar te maken in één omgeving.
EU	European Union	Hopelijk heb je hier wel eens van gehoord!
GDPR	Global Data Protection Regulation	Regelgeving die je voorschrijft hoe een bedrijf met persoonsgegevens moet omgaan.
GIT	Whatever rocks your boat, but it could be referring to Torvald (the creator) himself	Open source software voor gedistribueerd versiebeheer, in wezen een domme content tracker (zo zeggen ze zelf).
GKE	Google Kubernetes Engine	Google's versie van Kubernetes, beschikbaar gesteld als een beheerde omgeving.
GPU	Graphical Processing Unit	Het tweede brein van de computer. Bedoeld om het display aan te sturen, maar door de manier waarop deze processor werkt, is het erg handig om ook complexe berekeningen voor AI uit te voeren.
HR	Human Resources	Eigenlijk het personeelsbestand van een organisatie, hoewel we het meestal hebben over het management van het personeelsbestand.
I/O	Input and Output	De bits and bytes die uitgewisseld worden tussen systemen.
IaaS	Infrastructure as a Service	Het aanbieden of gebruiken van Infrastructuur als een dienst. Het is niet langer nodig om jouw eigen hardware aan te schaffen, maar deze kan op aanvraag gebruikt worden.
IBAN	International Bank Account number	Internationaal overeengekomen afspraak voor het identificeren van bankrekeningen over de nationale grenzen heen.

IBM	International Business Machines	Een geweldig bedrijf (althans dat vinden wij) met als strategie het ontwikkelen van AI- en cloudoplossingen.
IT	Infrastructure Technology	Met infrastructuur bedoelen we in deze context computerinfrastructuur zoals servers, opslag, netwerk, applicaties, besturingssystemen, etc.
ITIL	Information technology infrastructure library	Een set best practices voor IT-servicemanagement en IT-beheer om de beste services te leveren.
LDAP	Lightweight Directory Access Protocol	Een protocol dat toegang tot bronnen in een computernetwerk mogelijk maakt.
ML	Machine Learning	Een discipline van AI die tot doel heeft computers te leren leren en handelen zonder expliciet geprogrammeerd te zijn.
MVP	Minimum Viable Product	Een eerste versie van een product die gebruikt kan worden in de bedrijfsvoering, en dus gebruikt kan worden door klanten. Deze zal iteratief worden bijgewerkt met de verzamelde feedback.
OCR	Optical Character Recognition	Een technologie die tekst in een digitale afbeelding herkent. Het wordt vaak gebruikt om tekst in gescande documenten en afbeeldingen te herkennen.
OWASP	Open Web Application Security Project	Internationale non-profitorganisatie die zich toelegt op de beveiliging van webapplicaties. Het doel is om website-eigenaren en beveiligingsexperts te helpen webapplicaties te beschermen tegen cyberaanvallen.
PaaS	Platform as a Service	Het aanbieden of gebruiken van een volledig platform als een dienst. Dit omvat infrastructuur, een bundel gerelateerde software en de bijbehorende beheerdiensten. Voorbeelden hiervan zijn een ontwikkelplatformdienst zoals ARO (Azure Red Hat Openshift) maar het kan ook een dataplatform zijn.
PDF	Portable Document Format	Wij zijn er zeker van dat je bekend bent met het format.
PS	Post Script	Het komt uit het Latijnse postscriptum, dat schrijf je op het einde als je iets belangrijks vergeten bent en dat graag wil benadrukken. Je kunt ook PPS toevoegen als je nog iets bent vergeten. Maar wat moeten we zeggen over iemand die zoveel dingen begint te vergeten...
REST	Representational State Transfer	Een architectuur die een manier biedt om niet geïntegreerde applicaties te koppelen.
RISC	Reduced Instruction Set Computer	Verminder het aantal cycli per instructie. Het tegenovergestelde van CISC.

RPA	Robotic Process Automation	Automatiseringstechnologie die zorgt voor een digitale robot die repetitieve handmatige taken van ons kan overnemen waar we dagelijks gek van worden.
RPO	Recovery Point Objective	Maximaal acceptabele hoeveelheid data verlies.
RTO	Recovery Time Objective	Maximaal hoeveelheid tijd die gebruik mag worden om de systemen te herstellen.
SaaS	Software as a Service	Het leveren van software als een dienst.
SAP	Systems, Applications, Products	Een Duits software bedrijf.
SIEM	Security Information and Event Management	Technologie die de detectie van bedreigingen, compliance en het beheer van beveiligingsincidenten ondersteunt door informatie uit verschillende bronnen samen te voegen en te analyseren.
SoE	Systems of Engagement	Systemen (applicaties) die communiceren met klanten, medewerkers en partners.
Sol	Systems of Insight	Systemen die de data vertalen naar inzichten. Dit is waar dit boek over gaat.
SoR	Systems of Record	Een groep systemen die de kerngegevens van het bedrijf bevat. Dit kunnen klanten, burgers, producten, transacties, etc. zijn.
SQL	Structured Query Language	Het goede oude SQL is een programmeertaal om relationele databases te bevragen.
SRE	Site Reliability Engineer	Een rol voor iemand die een computerplatform beheert in plaats van één systeem. Dit concept is door Google geïntroduceerd om lineaire schaalvergroting van ondersteuningsteams te voorkomen.

OVER DE SCHRIJVERS

Jan Schravesande is een IBM en Open Group gecertificeerd Executive Enterprise Architect en heeft ervaring in verschillende bedrijfsdomeinen zoals Verzekeringen, Supply Chain, Energiesector en de Overheid. De laatste tien jaar werkt hij in de rol van strategisch adviseur voor Nederlandse overheid organisaties. Hij is regelmatig spreker op het gebied van data-gedreven architectuur en IT duurzaamheid.



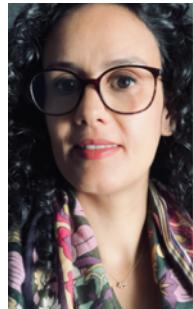
“Ik vind het fascinerend dat organisaties hun data gebruiken om bedrijfsresultaten te voorspellen en data een rol laten spelen bij advies over acties om de concurrentie voor te blijven of te excelleren in de beste dienstverlening. Data is zoveel meer geworden dan alleen maar een asset. Het is de bewaarder van alle kennis geworden, klaar om zijn geheimen te onthullen wanneer de juiste sleutel wordt gebruikt....”

Ronald Meijer is als gecertificeerd Senior IT Architect in dienst bij IBM en heeft ervaring met klanten in verschillende sectoren, zoals de financiële sector, transport en industrie. Tegenwoordig is hij als architect verantwoordelijk voor organisaties binnen de Nederlandse overheid. Ronald is een gepassioneerd docent op het gebied van TOGAF(R), Architecture Thinking en Microservices. Hij heeft in de loop van de jaren veel ervaring opgedaan in de diverse architectuur domeinen, van business tot infrastructuur architectuur. Hij is afgestudeerd elektrotechnicus en regelmatig spreker op diverse conferenties.



“Met de opkomst van AI vergeten we nog wel eens dat dit alleen goed kan werken als onze data op orde is. Daarnaast is er allerlei Internationale wet en regelgeving op het gebied van data en AI. Proportionaliteit en doelbinding zijn twee “principes” binnen de Nederlandse overheid waarmee we grenzen aangeven. De samenwerking tussen (AI gebaseerde) technologieën, processen, wet en regelgeving en verschillende culturen bepalen uiteindelijk het succes van de data-gedreven organisatie. Ik hou ervan om deze complexiteit te ontrafelen tot behapbare brokken”. “Dream big, Act realistic!”

Laila Fettah is een Open Group en IBM gecertificeerde Senior IT-specialist. Ze is begonnen in de wereld van de statistiek en heeft de eerste tien jaar van haar carrière besteed aan het vinden van wat er echt toe doet voor bedrijfsprocessen in gestructureerde en ongestructureerde data. Ze heeft als grote passie het bij elkaar brengen van mensen, processen en technologie. Momenteel is ze Architect en werkt daarbij aan overkoepelende thema's binnen de Nederlandse overheid.



“Dit heeft haar ertoegelicht om in haar huidige architectenrol deze twee holistisch samen te brengen, met een knipoog naar haar verleden: “make it simple, but significant”.

De informatie in dit boek is vrij te gebruiken zolang je als doel hebt om een data-gedreven organisatie te bouwen. Je kunt bij ons terecht met vragen, opmerkingen en we staan open voor aanvullingen en discussies.

Het boek is in PDF te downloaden van GitHub:
[https://github.com/DataDrivenBlueprint/
Downloads#readme](https://github.com/DataDrivenBlueprint/Downloads#readme)

Of gebruik de QR code:

Ook de poster vind je op GitHub



Laila Fettah: lailafettah@nl.ibm.com

Ronald Meijer: meijerr@nl.ibm.com

Jan Schravesande: schravesande@nl.ibm.com

Beste lezer,

Dit boek is bedoeld voor mensen die aan de slag willen met het bouwen van een data-gedreven organisatie. Ons doel is om een holistisch beeld te geven van de diverse capabilities die daarvoor nodig kunnen zijn. We denken dat we behoorlijk compleet zijn en we hebben gekozen voor om dit onderwerp in de breedte te behandelen omdat we geloven dat overzicht, samenhang en een breed begrip de belangrijkste startpunten zijn om een succesvolle data-gedreven organisatie te ontwikkelen.

Je krijgt te maken met veel stakeholders die je allemaal mee zult moeten nemen in jouw gedachtengoed. Dat is precies waar dit boek voor bedoeld is. Mensen die niet dagelijks hiermee te maken hebben of juist de mensen die met een specialistisch onderdeel bezig zijn, willen we helpen om een compleet overzicht te krijgen. Dit boek is geschreven door drie IBM-ers die geïnteresseerd zijn in dit onderwerp en in de loop van de tijd een passie hiervoor hebben ontwikkeld.

De kennis en inzichten die we in de loop van de tijd hebben verzameld door gesprekken met experts hebben we omgezet in een boek zodat we dit kunnen delen met iedereen die geïnteresseerd is in dit onderwerp.

Inmiddels is dit de vierde versie van dit boek. Het geeft wel aan dat het onderwerp in ontwikkeling is en ook de brede interesse voor dit onderwerp. We hopen dat je net zoveel plezier beleeft aan het lezen van dit boek als wij hadden met het schrijven.

Amsterdam,
December 2024

