STARTARCHITECTUUR GEGEVENSFUNDAMENT DATA GEDREVEN WERKEN

Datum 04 oktober2023

Status Definitief concept

Auteur Projectgroep gegevensfundament

Inhoudsopgave

1. Inleiding startarchitectuur gegevensfundament

1.1 Visie data gedreven werken DOWR-gemeenten

In 2021 hebben de DOWR-gemeenten een gezamenlijke visie ontwikkeld op data gedreven werken. In die visie zijn de doelen vastgelegd.

Dit zijn:

- **Het verbeteren van de dienstverlening voor onze inwoners en organisaties**
- **Het verbeteren (effectiever en efficiënter) van de bedrijfsvoering**
- **Het beter ondersteunen van de digitale samenleving.**

Deze doelen zijn vertaald naar een ambitie op het gebied van data gedreven werken. Alle 3 de organisaties willen meer volwassen worden op dit gebied.

Centraal in de aanpak om deze groei in volwassenheid te bereiken staan de 6 gedefinieerde bouwstenen:

1. Benutten van data
2. Organisatie en besturing
3. Mensen en competenties
4. Datakwaliteit
5. Architectuur
6. Informatiebeveiliging, privacy en ethiek

Voor iedere bouwsteen geldt dat binnen de kaders van deze bouwsteen stappen gezet moeten worden om toe te groeien naar een meer volwassen data-gedreven-organisatie(s).

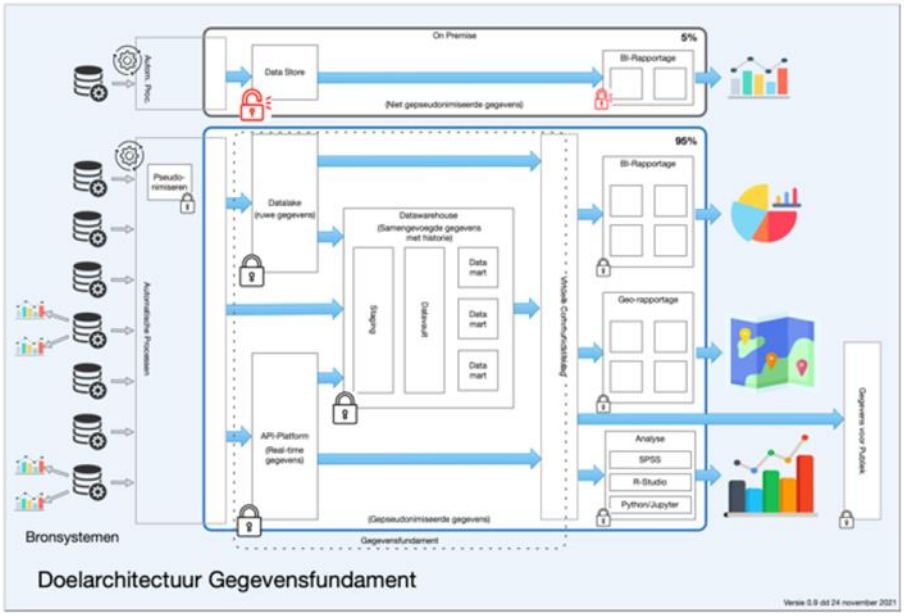
1.2 Bouwsteen architectuur

Een belangrijke randvoorwaarde om op een volwassen manier data gedreven te kunnen werken is de realisatie van een gegevensfundament. De belangrijkste argumenten om een gegevensfundament te realiseren zijn:

- Het voorziet in een centrale en betrouwbare vindplaats van data
- Het risico op fouten neemt af
- Het is wettelijk verplicht: een gegevensfundament borgt een veilige opslag van data en zorgt er voor dat voldaan kan worden aan de wettelijke bepalingen omtrent het opslaan van data (voorbeeld ter illustratie: voor data geldt ook een vernietigingstermijn, het fundament dwingt meta datering af, en daarmee is ook inzichtelijk wanneer het betreffende record moet worden vernietigd)
- Het maakt het realiseren van informatieproducten over de domeinen heen (meerdere bronnen) mogelijk
- Het is een effectieve en efficiënte manier om data op te slaan en te gebruiken
- Het maakt het actief sturen op de kwaliteit van data die binnen het fundament is opgeslagen mogelijk (bijvoorbeeld: eisen minimale datakwaliteit en automatische controle van data die ingebouwd kunnen worden)
- Het gegevensfundament biedt de mogelijkheid dat (privacy) gevoelige gegevens goed en veilig kunnen worden opgeslagen om te worden toegepast in informatieproducten (voorbeeld: vroeg signalering schuldenproblematiek)

Om deze redenen is in 2021, als vervolg op de vaststelling van de visie, een doelarchitectuur uitgewerkt. Deze doelarchitectuur is vastgesteld en gebruikt voor de onderbouwing van de budgetaanvraag. Dit advies is opgesteld door A. Brienen (2021).

Deze doelarchitectuur is de basis voor het voorliggende document, die de startarchitectuur van het gegevensfundament beschrijft.



1.3 Huidige situatie

In het architectuuradvies uit 2021 is al globaal ingegaan op de huidige situatie. Die laat zich kenmerken door:

- Lijstjes voeren de boventoon, relatief weinig bronnen zijn automatisch ontsloten
- Er worden veel informatieproducten naast elkaar ontwikkeld
- Er vindt geen centrale sturing plaats
- De basis onder deze informatieproducten (een betrouwbaar, veilig en centraal datamagazijn) is niet volledig op orde

Daarnaast is ook geconstateerd dat er grote verschillen zijn in het gebruik van data(producten). Zo is Deventer bijvoorbeeld, een van de eerste Nederlandse gemeenten die een Digital Twin heeft.

Over deze globale karakterisering van de huidige situatie is ten behoeve van de startarchitectuur een nadere verdieping gemaakt. In 3 aparte documenten is de huidige situatie beschreven.

1. Een inventarisatie van bestaande monitoren
2. Een globale architectuur van het GEO-domein (hoe wordt in het GEO-domein data gedreven gewerkt)
3. Een inventarisatie van de potentiële bronnen die binnen het fundament ontsloten kunnen worden

1.4 Kaders architectuur

Binnen DOWR-gemeenten zijn in het verleden architectuurkaders en -richtlijnen opgesteld. Op dit moment worden deze herijkt. De belangrijkste kaders die voor het gegevensfundament relevant zijn, zijn:

I. Informatieveiligheid en privacy op passend en vastgesteld niveau
II. Stabiliteit en continuïteit heeft prioriteit
III. Scheiding tussen applicaties en data
IV. Hergebruik gegevens
V. De DOWR-gemeenten bieden de inwoner een goede informatiepositie
VI. Informatie gestuurd werken met hulp van big, smart en open data
VII. Smart sourcen, uitgangspunt is Cloud tenzij[1]

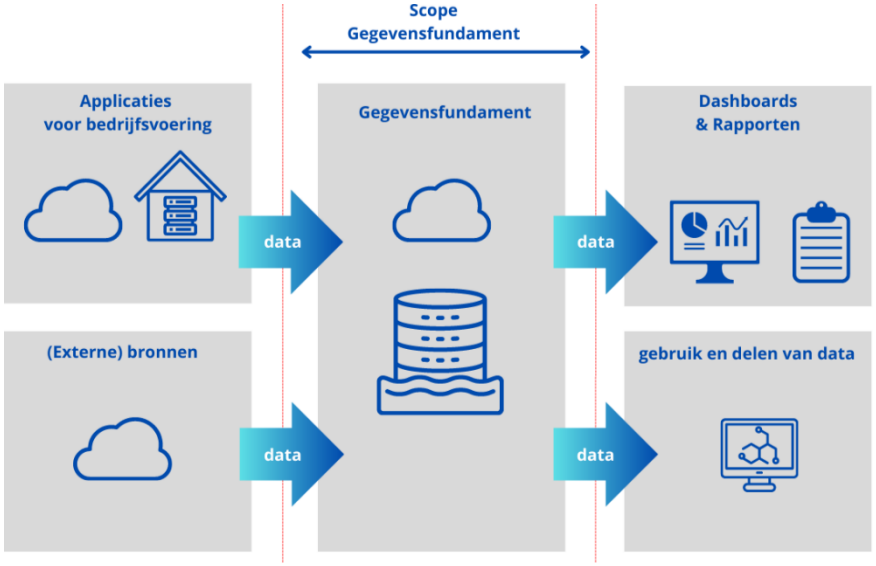
[1] Sluit aan bij het Rijksbreed Cloudbeleid 2022

VIII. Gegevensuitwisseling via openstandaarden
IX. Volledig en sluitend Informatiebeheer[2]

De keuzes en oplossingsrichting die in dit document worden gepresenteerd zijn allen getoetst aan deze richtlijnen.

1.5 Scope

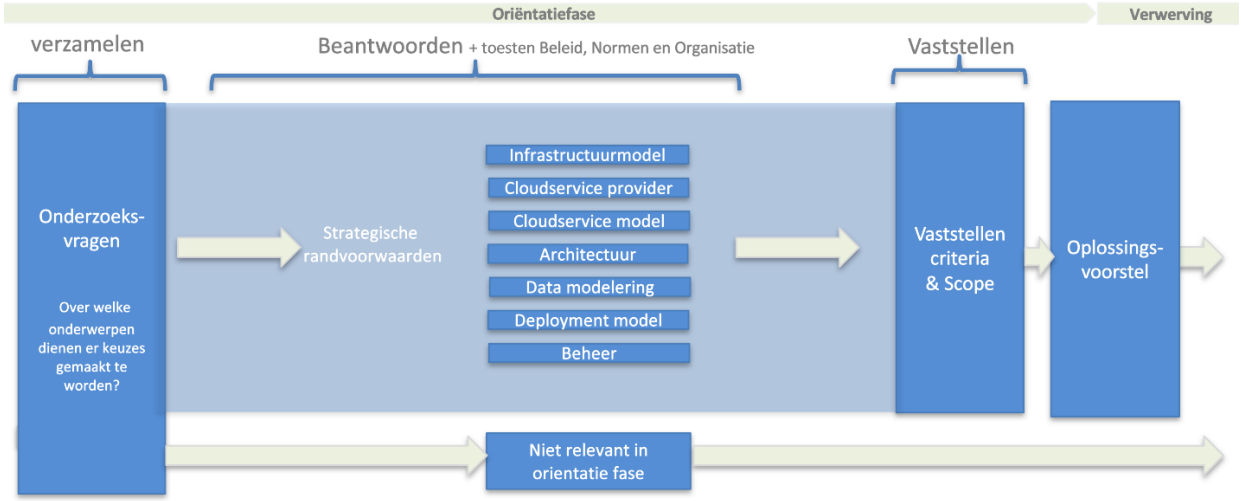
Het gegevensfundament onttrekt de data uit verschillende bronnen welke afkomstig zijn uit verschillende bedrijfsvoering applicaties en andere externe bronnen. Hoewel deze applicaties dienen als bron, vallen ze niet binnen de scope van het oplossingsvoorstel van het gegevensfundament. Het gegevensfundament dient als basis waarop rapporten en dashboards kunnen worden ontwikkeld. Het ontwikkelen van deze rapporten en dashboards valt echter ook buiten de scope van het oplossingsvoorstel van het gegevensfundament. De scope van het oplossingsvoorstel omvat de beschrijving van de techniek van het gegevensfundament.



Om deze globale beschrijving van de scope van het fundament wat nader te duiden en onduidelijkheden weg te nemen is een document gemaakt met fictieve userstories. Aan de hand van deze concrete gebruikersvragen zal duidelijk worden wat het wel en niet behelst.

1.6 Verantwoording

In onderstaande figuur is het proces geschetst wat doorlopen is om te komen tot dit uiteindelijke voorstel.



[2] Voor data geldt ook een vernietigingsplicht

1.7 Leeswijzer

Hoofdstuk 3 beschrijft de startarchitectuur voor het gegevensfundament dat voor de DOWR-gemeenten zal worden gerealiseerd. De startarchitectuur wordt beschreven op een 3-tal niveaus:

1. Componenten (paragraaf 3.1)
2. Doelarchitectuur (paragraaf 3.2)
3. Startarchitectuur (paragraaf 3.3)

De startarchitectuur is gebaseerd op een aantal strategische uitgangspunten die in juli 2023 door de stuurgroep gegevensfundament zijn vastgesteld. Deze keuzes rondom de strategische randvoorwaarden worden in hoofdstuk 2 beschreven en beargumenteerd. Welke keuzes dit betreft is opgenomen in de figuur in de vorige sub paragraaf.

2. Strategische Randvoorwaarden

In de volgende paragrafen worden de strategische voorwaarden in de volgende logische volgorde uitgewerkt. Van Infrastructuurmodel, CloudService Provider en Cloud Service model via Architectuur en datamodellen naar beheer.

2.1 Infrastructuurmodel

Het Infrastructuurmodel verwijst naar de fysieke en virtuele componenten die worden gebruikt om een gegevensfundament te bouwen in verschillende IT-omgevingen, zoals On-Premise (lokaal datacenter), Hybride Cloud (combinatie van On-Premise en Cloud), of volledige Cloud-omgeving.

2.1.1 Opties

Er zijn een drietal mogelijkheden:

I. Cloud
II. On-Premise
III. Hybride

Belangrijke elementen bij de keuze voor een infrastructuurmodel zijn flexibiliteit & schaalbaarheid, toekomstvastheid, beveiliging & privacy, continuïteit en kosten.

Onderdeel Cloud heeft weer twee keuzemogelijkheden

I. Private Cloud
II. Public Cloud

Een private cloud kent echter op voorhand al zoveel (financiële) nadelen dat deze optie voor dit onderzoek buiten beschouwing is gelaten. Als gesproken wordt over de cloud wordt dus gesproken over een public-cloud oplossing.

2.1.2 Voordelen/Nadelen

	Optie 1: Cloud	Optie 2: On-Premisse	Optie 3: Hybride
Beleid	Past bij de ambitie zoveel mogelijk naar de Cloud te brengen.	Past binnen het beleid ‘Cloud, tenzij’ wanneer er gegronde redenen zijn om hiervan af te wijken.	
Kosten	Betaalt naar gebruik, geen investering. Kosten-efficiënt, als best practices worden gebruikt en governance goed is ingericht.	Initiële investering, onderhoud, beheer	Zowel investering als abonnement. Duurste oplossing
Innovatie	Snelle adoptie van innovatie, direct beschikbaar	Conservatief. Technologie van gister. Weinig doorontwikkeling	Combinatie van beide. Zowel innovatief en conservatief
Beslag op organisatie	Veel werk gedaan door Cloudprovider.	Legt groot beslag op resources uit de organisatie	Legt groot beslag op resources uit de organisatie
Tijdlijn	Turnkey	Lange opbouwtijd	Lange opbouwtijd met turnkey elementen.

Veiligheid & Privacy	Past met de juiste maatregelen binnen het privacy- en beveiligingsbeleid van DOWR-i. State of the Art		Past met de juiste maatregelen binnen het privacy- en beveiligingsbeleid van DOWR-i. Complex en Tijdrovend		Slechtst van beide werelden vanwege koppelvlak Cloud/On-Premise.	
Flexibiliteit & Schaalbaarheid	Gebruik en betaal voor wat nodig is. Op- en afschalen on-demand. Productkeuze beperkt(er), leverancier specifiek.		Benodigde hardware beperkende factor. Opschalen kost tijd. Afschalen niet mogelijk. Keuzevrijheid hoog voor hardware en software. Open source mogelijk		Beste van beide werelden. Op- en afschalen en ook keuzevrijheid maximaal	
Toekomst	Flexibeler, innovatiever, trend.		Star, weinig innovatie, geen doorontwikkeling		Starre onderdelen gecombineerd met flexibele innovaties	
Continuïteit	Eenvoudig		Complex. Veel scenario's om rekening mee te houden		Basis complex, veel scenario's. Cloud gedeelte makkelijk te continueren.	

Verschillen Private Cloud en Public Cloud

De Public Cloud wordt gedefinieerd als computingservices die door externe providers worden aangeboden via het openbare internet, zodat ze beschikbaar zijn voor iedereen die ze wil gebruiken of kopen. Ze kunnen gratis zijn of te koop op aanvraag, waardoor klanten betalen voor de gebruikte CPU-cycli, opslag of bandbreedte.

In tegenstelling tot een Private Cloud kan met een Public Cloud worden bespaard op kosten voor de aanschaf, het beheer en het onderhoud van de on-premises infrastructuur voor hardware en toepassingen.

2.1.3 Advies

POSITIONEER HET GEGEVENSFUNDAMENT VOLLEDIG IN DE (PUBLIC) CLOUD

Omdat:

- Het data gedreven werken voor DOWR nieuw is, daarmee is het voorspellen van behoeften en benodigdheden risicovol. Een Public Cloud oplossing is hierbij de beste oplossing omdat:
  - Cloud oplossingen maximaal flexibel en schaalbaar zijn
  - Cloud biedt de schaalbaarheid voor groeiende datavolumes
  - Cloud faciliteert 'learning by doing' voor een lerende organisatie
- Een Cloud oplossing vraagt niet om een initiële (eenmalige) investering. Kosten zijn gerelateerd aan het gebruik. Daarmee is een Cloud oplossing een kosten efficiënte oplossing mits de governance goed wordt ingericht en best-practices worden gebruikt.
- Het aansluit op het Rijksbreed Cloudbeleid 2022 en de aanzet gemeentelijk Cloud beleid 2023.
- Cloud ondersteunt veilig werken met gevoelige gegevens door het aanbieden van uitbreidbare beveiligingsmaatregelen

De keuze voor een volledige Cloud oplossing wijkt af van de referentiearchitectuur uit 2021. Het belangrijkste argument uit 2021 voor deze oplossing was dat dit de meest veilige oplossing zou zijn om de meest gevoelige gegevens op te slaan. Inmiddels kan ook binnen een Cloud omgeving een net zo hoog beschermingsniveau worden gerealiseerd als bij lokale opslag.

2.2 Cloud Serviceprovider

Een CloudService-Provider (CSP) is een bedrijf dat Cloud computing-diensten aanbiedt aan individuen, organisaties en andere organisaties. Cloud computing omvat het leveren van computerresources zoals rekenkracht, opslag, databases, software en andere services via het internet.

2.2.1 Opties

Er zijn 4 relevante opties om uit te kiezen:

I. Microsoft Azure (Azure)
II. Amazon Web Services (AWS)
III. Google Cloud Platform (GCP)
IV. Multi-Cloud (Een combinatie van meerdere CSP's)

Alhoewel er meer CloudService Provider opties zijn valt het overgrote merendeel af en zijn deze op dit moment niet relevant om verder te behandelen. Dit is o.a. vanwege de volgende redenen:

- Aanbod van diensten niet voldoende voor volledig gegevensfundament
- Kleinere schaal en bereik
- Minder uitgebreide dienstenportfolio, onvoldoende voor gegevensfundament
- Minder bekendheid en adoptie (volwassenheid)
- Mogelijke zorgen over locatie van data in instabiele en/of (geo)-politiek risicovolle regio's[3]

2.2.2 Voordelen/Nadelen

De eerste drie CSP's zijn qua functionaliteit, betrouwbaarheid, volwassenheid en kosten ongeveer gelijk. Nummer vier in de rij, de Multi-Cloud oplossing is complexer en biedt momenteel geen voordelen ten opzichte van de andere opties.

Pluspunten van Microsoft Azure boven de andere opties zijn:

- DOWR-gemeenten zijn al aan Microsoft gelieerd
  - Integratie met bestaande Office365, SharePoint en PowerBI omgeving mogelijk
  - Kennis en ervaring Microsoft omgevingen aanwezig
- Meeste overheidsinstantie gebruiken Microsoft Azure. Past ook in Cloudbeleid Rijksoverheid
- Leer omgeving (zandbak) snel en veilig op te zetten
- Van de drie CSP's legt Microsoft meeste nadruk op Beveiliging en Compliance

2.2.3 Advies

BOUW HET GEGEVENSFUNDAMENT OP IN DE AZURE TENNANT

Omdat

- DOWR-I voor Microsoft-producten kiest tenzij andere producten aantoonbaar beter in de behoeften voorzien
- Voor de CSP is niet aannemelijk gemaakt dat een andere CSP een betere oplossing is
- Integratie met andere platforms eenvoudiger zal zijn omdat deze ook allemaal uit de Microsoft-stal komen (bijvoorbeeld: authenticatie via de Microsoft Authenticator App)

[3] Vanuit de AVG is een belangrijke randvoorwaarde dat persoonsgegevens alleen mogen worden gebruikt of opgeslagen in landen die net zo veilig zijn als de Europese Unie zelf voorschrijft. Dat maakt werken in veel landen (met name de Verenigde Staten) erg lastig

2.3 Deploymentmodel

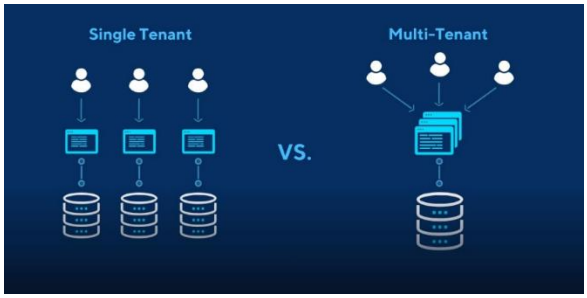
Het Deployment model verwijst naar de manier waarop de toegang tot de infrastructuur wordt ingedeeld.

2.3.1 Opties

Voor de deployment zijn 2 opties mogelijk:

I. Single-Tenant
II. Multi-Tenant

In de figuur zijn beide opties grafisch weergegeven.



Ad I. Single-Tenant

In een Single-tenant Deployment model heeft elke gemeente zijn eigen afzonderlijke infrastructuur, database en toepassingslogica, waardoor volledige isolatie tussen gemeenten wordt bereikt.

Ad 2: Multi-Tenant

Bij een Multi-tenant is die scheiding tussen de drie gemeenten er ook, alleen deze scheiding is virtueel. De DOWR-gemeenten delen dezelfde infrastructuur, database en toepassingslogica, maar de gegevens van elke gemeente worden strikt gescheiden en geïsoleerd. Hierdoor lijkt het voor elke gemeente alsof ze een eigen speciale instantie van de software hebben, terwijl ze in werkelijkheid gebruikmaken van dezelfde gedeelde bronnen.

2.3.2 Voordelen/Nadelen

	Single-Tenant	Multi-Tenant
Kosten	3x hoger qua infrastructuur en beheer(s)kosten	Eenmalige infrastructuur- en beheerskosten (tijd)
Schaalbaarheid	Extra tenant nodig	Nieuwe gemeente(n) eenvoudig toe te voegen
Data isolatie	Fysiek	Virtueel
Piekbelasting	Eigen resources	Performanceproblemen zeldzaam in Cloud
Eigen configuratie per organisatie	Mogelijk	Niet mogelijk, gedeelde infrastructuur
Beveiliging	Door fysieke scheiding extra beveiliging	Een Fysieke scheiding van data is beter dan een virtuele. Technisch mogelijk om gegevens uit andere omgeving te zien, indien niet juist ingericht

DOWR heeft geen juridische verplichting om data in een fysiek afzonderlijke infrastructuur te plaatsen. Het belangrijkste aandachtspunt blijkt te zijn dat gemeente-specifieke data op de juiste wijze virtueel geïsoleerd wordt opgeslagen wanneer dit nodig is, terwijl generieke en niet-privacygevoelige data gedeeld kan en mag worden.

2.3.3 Advies

RICHT HET GEGEVENSFUNDAMENT IN ALS MULTI-TENANT

Omdat:

- Dit het beheer van de omgeving eenvoudiger en goedkoper maakt
- Dit binnen DOWR ook de meest gangbare manier van werken is
- Er ook binnen DOWR maar één Microsoft-tenant is (multi tenant dus)
- Er geen wettelijke verplichting is om data fysiek te scheiden
- Via metadatering en andere manieren voldoende mogelijkheden zijn om gevoelige data op een goede manier virtueel te scheiden
- Dit het uitwisselen van niet gevoelige data tussen de DOWR-gemeenten eenvoudiger maakt

2.4 CloudService model (IaaS, PaaS en/of SaaS)

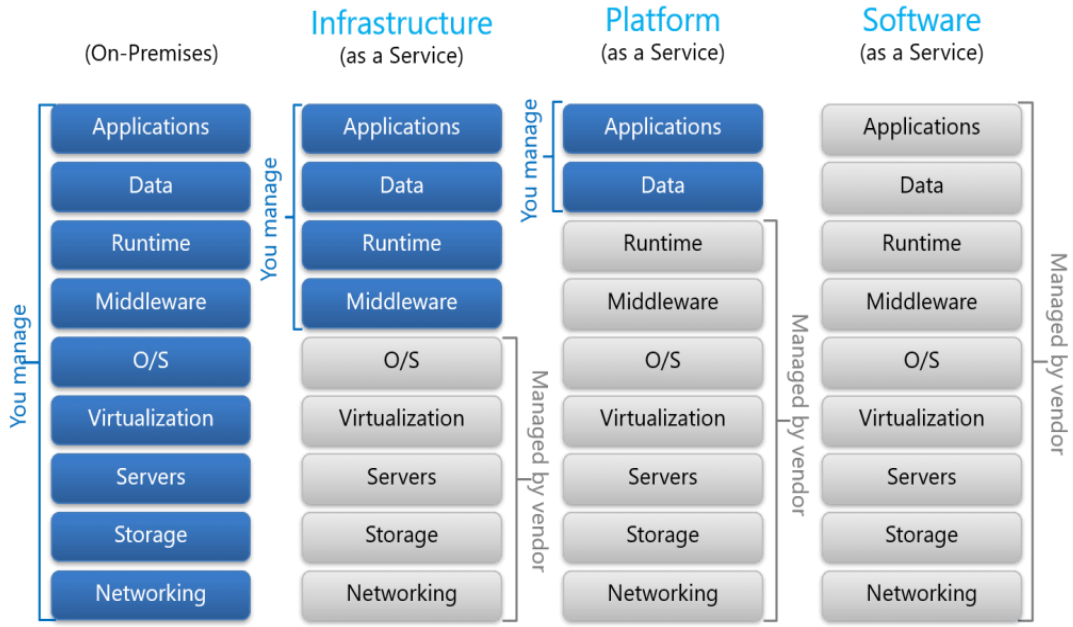
Een CloudService model verwijst naar de verschillende manieren waarop Cloud computing-diensten worden geleverd aan gebruikers en organisaties. Deze modellen definiëren de gedeelde verantwoordelijkheden van zowel de Cloudprovider als de gebruiker (DOWR) met betrekking tot beheer, onderhoud en toegang tot Cloud resources. Ongeacht het gekozen CloudService zijn er taken die door DOWR gedaan moet worden. Toegangsbeheer, ETL van de gegevens zelf en het opzetten van de accounts zijn slecht enkele voorbeelden van deze activiteiten.

2.4.1 Opties

Er zijn drie keuzes

I. Infrastructure as a Service (IaaS):
II. Platform as a Service (PaaS):
III. Software as a Service (SaaS):

De onderstaande figuur beschrijft deze varianten.



Wellicht ten overvloede: De blauwe vlakken geven aan welke onderdelen DOWR zelf zal moeten ‘bemensen’. De grijze vlakken worden door de CSP uitgevoerd.

2.4.2 Voordelen/Nadelen

Een IaaS inrichting is in hoofdlijnen het meest vergelijkbaar met een On-Premise inrichting. Veelal op basis van Virtual Machines worden oplossingen gehost waarbij de gebruiker zelf verantwoordelijk is voor updates, inrichtingen, monitoring van virtual machines, databases, etc. Deze optie valt af vanwege het ontbreken van Cloud voordelen zoals schaling, onderhoud, en het gebruik kunnen maken van innovatieve diensten. In de praktijk worden IaaS inrichtingen voornamelijk gebruikt door bedrijven die zelf software ontwikkelen en hosten.

Platform as a Service (PaaS) staat voor "Platform als een Dienst". Het is een Cloud computing model dat een platform aanbiedt voor het ontwikkelen, implementeren en beheren van applicaties zonder dat men zich geen zorgen hoeft te maken over de onderliggende infrastructuur. PaaS gaat verder dan IaaS omdat het niet alleen de infrastructuur levert, maar ook een platform biedt voor applicatieontwikkeling en -uitvoering.

De grotere CSP's hebben een uitgebreid portfolio aan modulaire bouwblokken beschikbaar via een PaaS platform, dit stelt de gebruiker in staat om voor een specifieke wens of situatie een maatwerk oplossingsrichting op te tuigen.

Bij Software as a Service (SaaS) wordt ook de applicatie als service geleverd. De CSP is verantwoordelijk voor onderhoud, updates en patches van de software. Eindgebruikers hebben altijd toegang tot de laatste versie van de software. DOWR hoeft zich dus voornamelijk te richten op de gebruikers, de data en de applicatie zelf, helaas met beperkte configuratiemogelijkheden.

Het uitgangspunt bij SaaS-diensten is dat de gebruiker de koers en richting van de dienst omarmt, dit betekent enerzijds een beperkte invloed op configuratie en inrichting, anderzijds dat marktontwikkelingen en nieuwe technieken vanuit de CSP worden geleverd.

Het onderscheidt tussen SaaS en PaaS is beschreven in onderstaande tabel.

	PaaS		SaaS	
Applicatie	Maatwerkapplicaties, Ontwikkelingsplatform, tools voor softwarebouw		Kant en Klaar, Beperkte configuratie	
Snelheid implementeren	Na de keuze, eerst programmeren dan implementeren.		Na de keuze, snelle implementatie	
Flexibiliteit	Eenvoudige migratie naar andere oplossing. Keuze voor applicatie minder cruciaal.		Laag. Keuze voor applicatie eenmaal gemaakt, dan lastig te veranderen naar PaaS of andere SaaS-oplossing.	
Benodigde Expertise	Veel voor programmeren en beheren		Weinig voor beheren	

2.4.3 Advies

Bouw het Gegevensfundament op met SaaS-producten waar mogelijk, PaaS waar nodig

Omdat:

- DOWR-i hanteert al min of meer een "SaaS-tenzij" beleid, zij het nog niet volledig gedocumenteerd of volledig in praktijk gebracht.
- Minder (technisch) beheer past in de strategie van DOWR-i

In Q2 2023 heeft Microsoft een nieuwe SaaS-dienst onder de naam Microsoft Fabric in public preview beschikbaar gesteld. Deze oplossing is gepositioneerd als volledige SaaS invulling van een gegevensfundament. Het is niet te verwachten dat de roadmap van Fabric aansluit bij de tijdlijnen van het gegevensfundament, maar de ontwikkelingen kunnen een interessante oplossing zijn voor de lange termijn. In gesprekken met Microsoft is vastgesteld dat de gekozen architectuur richting (zie hoofdstuk 3) aansluit bij de visie van Microsoft Fabric wat een eventuele migratie in de toekomst mogelijk maakt.

2.5 Datamodel

Een goed ontworpen datamodel is de basis voor het effectief en efficient opslaan van data van een organisatie. Daarnaast ondersteunt een goed datamodel het beheer van de opgeslagen data zodat voldaan kan worden aan wet- en regelgeving.

2.5.1 Opties

Er zijn twee keuzes:

I. Implementeren Gemeentelijke gegevensmodel (GGM)
II. Ontwikkelen eigen datamodel

Het GGM biedt een datamodel dat als standaard raamwerk kan dienen voor DOWR. Het GGM wordt toegepast binnen een aantal gemeenten en is ook ontwikkeld door gemeenten. In de meest recente editie van de GEMMA is het bedrijfsobjectenmodel gewijzigd en dient het GGM als basis.

De keuze tussen een eigen datamodel en het GGM hangt af van verschillende factoren, zoals de mate van standaardisatie binnen de sector, de complexiteit van de organisatieprocessen en de benodigde flexibiliteit.

2.5.2 Voordelen/Nadelen

	GGM	Eigen Datamodel
Standaardisatie	Hoog. Volledig volgens de basisregistraties opgebouwd	Laag. Maatwerk. Kennis en model zelf
Uitwisseling	Eenvoudig met andere gemeenten met GGM	Een uniek datamodel kan de samenwerking en gegevensuitwisseling met andere organisaties bemoeilijken.
Efficiëntie	Bestaand model. Weinig ontwikkel en implementatietijd	Het ontwikkelen en onderhouden van een aangepast datamodel vergt inzet van tijd en expertise.
Flexibiliteit	Door de hoge standaardisatie is flexibiliteit kleiner. Als gemeenten voor je echter wel zelf de regie over aanpassingen waardoor snelheid van aanpassingen alsnog vrij hoog kan zijn	Met de juiste kennis zijn aanpassingen snel te realiseren
Complexiteit	Het doorgronden van de onderdelen van het GGM en de onderlinge samenhang is complex. Niettemin kan het GGM stap-voor-stap ontdekt en gevuld worden.	Minimaal. Bevat alleen attributen die relevant zijn voor de organisatie

2.5.3 Advies

MODELLER DATA IN HET GEGEVENSFUNDAMENT OP BASIS VAN HET GGM

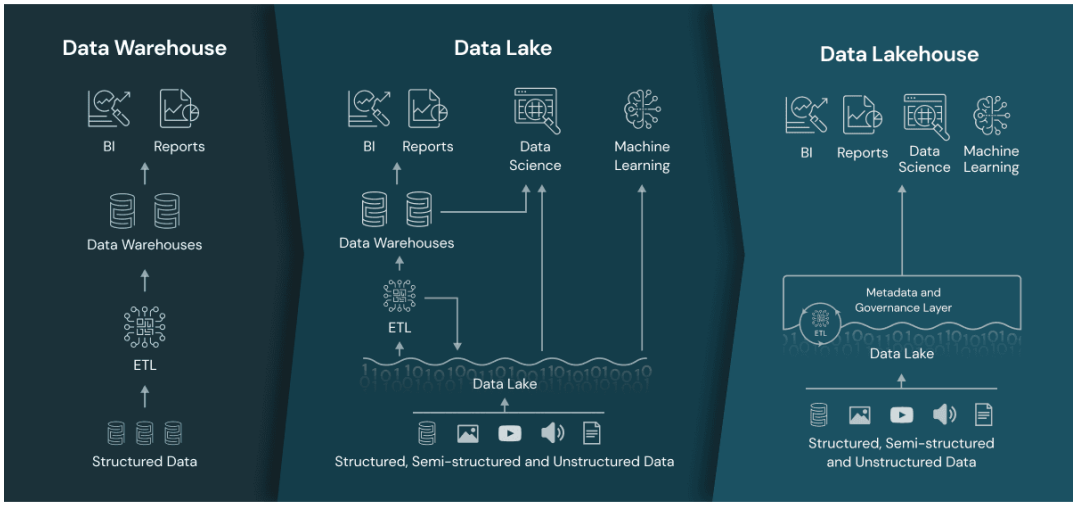
Omdat:

- De DOWR-gemeenten weinig expertise en ervaring hebben op het gebied van datamodellering
- Het GGM ontwikkeld is door en voor gemeenten (op basis van best practices)
- Het GGM inmiddels opgenomen is in de GEMMA architectuur *en* de VNG zich gecommitteerd heeft aan de verdere doorontwikkeling ervan
- Het GMM het delen van data met andere gemeenten aanzienlijk vereenvoudigt
- Het GGM continue wordt doorontwikkeld en verbeterd door gemeenten zelf (inmiddels onder regie van VNG)
- Het GGM past volledig in de Common Ground gedachte
- Het GGM voorziet in de registratie van ALLE basisregistraties inclusief het GEO-Domein

2.6 Type Architectuur

Het type architectuur legt vast hoe gegevens worden opgeslagen, beheerd, verwerkt, en geanalyseerd om van gegevens naar waardevolle inzichten te komen waarmee men data gedreven kan werken.

2.6.1 Opties



De figuur hierboven geeft de opties weer, dit zijn:

I. Data Warehouse architectuur
II. Data Lake architectuur
III. Data Lakehouse architectuur

Ad I: Data Warehouse Architectuur

Dit omvat een architectuur waarin de Data Warehouse centraal staat in het gegevensfundament. De opslag, verwerking, en integratie van de gegevens vindt allemaal hier plaats. De analyses en rapportages lezen vervolgens de gegevens uit de Data Warehouse.

Ad II: Data Lake Architectuur

Een Data Lake is een opslagplaats voor ruwe, ongestructureerde en gestructureerde gegevens. Deze architectuur omvat het opzetten van een Data Lake als centrale vindplaats voor alle gegevens.

Ad III: Data Lakehouse Architectuur

Een Data Lakehouse is een combinatie van een Data Lake en een Data Warehouse. Gegevens worden opgeslagen in een Data Lake en vervolgens verder bewerkt met Data Warehouse capaciteiten waar nodig. Het is, gelet op de beide andere modellen, een relatief nieuw, maar al wel volwassen concept.

2.6.2 Voordelen/Nadelen

In de tabel hieronder is een overzicht gegeven van de verschillen tussen het Data Warehouse en het Data Lake. Die verschillen geven meteen ook inzicht in de belangrijkste voor*- en nadelen van ieder model

	Data Warehouse		Data Lake	
Gestructureerde data	Uitermate geschikt		Uitermate geschikt	
Ongestructureerde data (GEO Domein)	Niet geschikt, beperking tot SQL		Uitermate geschikt	
Streaming data verwerken	Niet ideaal		Uitermate geschikt	
Garantie Datakwaliteit	ACID-regels dus garantie op datakwaliteit		Garantie niet mogelijk	
Analyse van data	Vaak extern met andere programmeertalen zoals Python		Gebruiksvriendelijk	
Kosten	Hoog		Efficiënt	
Toegankelijkheid	Vaak alleen ontwikkelaars en niet voor business.		Eenvoudig om 'tijdelijk' toegang te geven aan derden. Maakt (toegangs)beheer complexer	
Schaalbaar	Mogelijk		Hoog	

Data Lake House nader uiteengezet

Het Data Lakehouse principe is een relatief nieuw data-architectuur concept; het combineert elementen van Data Lakes en Data Warehouses om zo te komen tot het beste van beide werelden

Een Lakehouse maakt gebruik van de schaalbare en gedistribueerde opslag van een Data Lake om alle soorten data op te slaan; tabellen, Excel bestanden, video's, geluidsfragmenten etc. zijn allemaal op te slaan op een schaalbare manier.

Dit maakt het eenvoudiger om gegevens van verschillende bronnen in hun ruwe vorm te bewaren, zonder de noodzaak om deze gegevens eerst te transformeren voordat ze worden opgeslagen (zoals bij een Data Warehouse gebeurt).

Een van de belangrijkste voordelen van het Data Lakehouse principe is de flexibiliteit die het biedt. Het kan gemakkelijk worden aangepast en uitgebreid om nieuwe gegevensbronnen toe te voegen en de gegevensverwerking te verbeteren.

De zwaktes van de Lakehouse zijn eigenlijk vooral de mindere mate van ondersteuning van diensten en de mindere mate van ervaring die organisaties ermee hebben.

2.6.3 Advies

GEEF HET GEGEVENSFUNDAMENT VORM o.b.v. EEN DATALAKEHOUSE ARCHITECTUUR

Omdat:

- Het combineert het beste van beide bestaande architectuurtypes
- Het beter in staat is om te gaan met alle types data (bijvoorbeeld sensordata)
- Het kosten efficiënter is, niet alle data wordt opgeslagen in een duurder datawarehouse
- Hoewel het Lakehouse-principe nog relatief nieuw is en minder ondersteuning heeft dan Data Lakes en Data Warehouse s, zal de implementatie ervan bij DOWR in de loop der tijd meer functionaliteiten toevoegen.
- Er veel innovatie plaatsvindt op het relatief nieuwe Lake House concept, het is de basis voor het nieuwe Microsoft Fabric concept. Een migratie naar dit SaaS gegevensfundament behoort dan zeker tot de mogelijkheden in de toekomst.

Het feit dat een Data Lakehouse een relatief nieuw, maar al wel volwassen concept is, mag niet onvermeld blijven. Documentatie, ervaring en ondersteuning lopen daardoor achter op het veel oudere datawarehouse concept omdat:

a) Het meer mogelijkheden biedt
b) Het multi-hop-principe ondersteunt (zie paragraaf 3.2 waar hier nader op in wordt gegaan)
c) Het een evt. switch naar een SaaS-variant van het gegevensfundament (MS Fabric) eenvoudiger maakt

2.7 Beheer

Dit hoofdstuk focust zich op 2 kernvragen:
1. Welke rollen zijn er nodig om het gegevensfundament te beheren?
2. Hoe worden de benodigde rollen ingevuld?

2.7.1 Benodigde rollen

Bij de start van het fundament moeten in ieder geval de rollen zijn ingevuld die opgenomen zijn in de tabel hieronder.

Groep	Rol	Omschrijving
Toetsend	Information Security Officer	Ik zie toe op de naleving van het beleid op het gebied van informatieveiligheid/security.
	Privacy Officer	Ik zie toe op de naleving van het privacy beleid van de gemeente (DPIA)
Beheer	Technisch beheer	Ik zie toe op het, in technische zin, goed functioneren van de applicatie en de infrastructuur waarop deze draait
	Functioneel (applicatie) beheer	Ik zie toe op het, in functionele zin, goed functioneren van de applicatie en lever een bijdrage aan het verbeteren van het functioneren hiervan. Daarnaast ondersteun ik de gebruikers in het werken met de applicatie.
	Engineer gegevensfundament[4][5]	Ik zie erop toe dat de data in het fundament duurzaam wordt opgeslagen. Ik zorg ervoor dat we blijven voldoen aan wet- en regelgeving op het gebied van archiveren en vernietigen van data.

Mogelijk dat er na verloop van tijd, als data gedreven werken op grote schaal wordt toegepast nog extra rollen worden gevraagd. Deze zijn opgenomen in de tabel op de volgende pagina.

[4] Deze rol levert ook een wezenlijke bijdrage in het ontwikkelproces van nieuwe monitoren, door data, afkomstig uit de bron, zodanig te modelleren dat deze opgeslagen kan worden in het fundament
[5] In de praktijk is het onderscheid met de data engineer niet altijd even duidelijk, er zal overlap zijn tussen deze rollen

Groep	Rol	Omschrijving
Toetsend	Ethics Officer	Ik bewaak dat we blijven voldoen aan de opgestelde ethische kaders.
	Cloud Architect	Ik zorg ervoor dat de oplossing blijft voldoen aan de architectuur, het beleid en de principes op het gebied van werken in de (Azure) Cloud
	Data architect	Ik ervoor dat de data en de datamodellering in het fundament blijvend voldoet aan de eisen die hieraan gesteld worden vanuit zowel IT, governance als business perspectief

In de bijlage is per rol een globaal overzicht gegeven van de taken behorend bij deze rollen.

2.7.2 Invulling van de rollen

In deze paragraaf wordt ingegaan of de genoemde rollen wel/niet structureel een plek moeten krijgen binnen de DOWR-gemeenten. De tabel hieronder geeft hierover duidelijkheid. **Het gaat hierbij dus over de situatie dat het gegevensfundament al in beheer is genomen**.

Rol	Positionering	Reden
Information Security Officer	Intern (DOWR)	Dit een bestaande rol is die al intern is belegd en er vanuit het fundament geen reden is om dit te wijzigen
Privacy Officer	Intern (DOWR)	Dit een bestaande rol is die al intern is belegd vanuit het fundament geen reden is om dit te wijzigen
Technisch beheer	Extern	Dit onlosmakelijk verbonden met de keuze voor het realiseren van het fundament in de Cloud i.c.m. SaaS/PaaS Het technische beheer voor een cloudapplicatie wordt altijd uitgevoerd door de leverancier
Functioneel (applicatie) beheer	Intern (DOWR-I)	De meest belangrijke taak bestaat uit het in verbinding komen en blijven met de gebruikers. Om die reden is het op afstand uitvoeren van deze taken niet mogelijk.
Engineer gegevensfundament	Intern (DOWR-I)	Dit een rol is die om samenwerking vraagt met veel interne disciplines. Juist ook omdat er vanuit deze rol ook een bijdrage wordt geleverd aan de ontwikkeling van informatieproducten.

2.7.3 Advies

In de voorgaande paragrafen is aangegeven welke rollen noodzakelijk zijn en waar deze het beste gepositioneerd kunnen worden. Een korte samenvatting is weergegeven in de hieronder weergegeven tabel.

Het advies is verder om:

- Budgettair al rekening te houden met de inzet van een Ethics Officer omdat de ontwikkeling van Artificial Intelligence/ Machine Learning exponentieel groeit, en juist daar de noodzaak aanwezig is van ethische kaders en toezicht hierop;
- De realisatie/ opbouw van het gegevensfundament en de migratie van bestaande monitoren naar dat fundament deels te doen met een dubbele bezetting zodat de interne collega kan leren van de externe inhuur;
- Voor de externe inhuur in te zetten op het senior niveau.

Rol	Positionering
Information Security Officer	Intern
Privacy Officer	Intern
Technisch beheer	Extern
Functioneel (applicatie) beheer	Intern
Engineer gegevensfundament	Intern

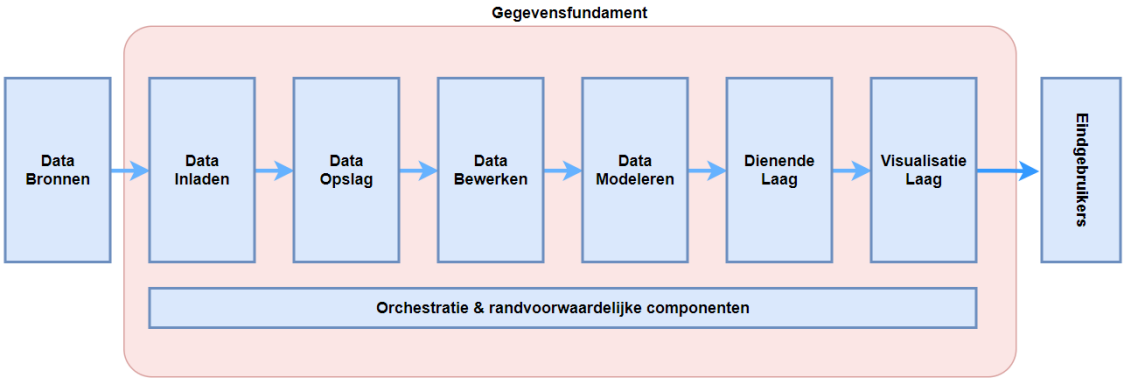
3. Doelarchitectuur/ Startarchitectuur

Dit hoofdstuk beschrijft de doelarchitectuur van het gegevensfundament. Gezien de fase van het traject en de verschillende doelgroepen is er gekozen om de beschrijving in drie verschillende abstractieniveaus te definiëren:

1. Een weergave op hoofdcomponenten; hierin zijn de architectuur bouwblokken die het gegevensfundament vormgeven weergegeven;
2. Doelarchitectuur; een verdiepingsslag waarin de componenten zijn uitgewerkt in specifieke diensten en functionaliteiten;
3. Startarchitectuur; waarin de specifieke diensten en functionaliteiten zijn vertaald naar mogelijke oplossingsrichtingen in het Azure platform.

Hierbij dient opgemerkt te worden dat de startarchitectuur een mogelijke oplossingsrichting beschrijft die niet limitatief is, maar vooral gericht is op vervolgstappen/aanbesteding.

3.1 Componenten



Figuur 1: Algemene componenten

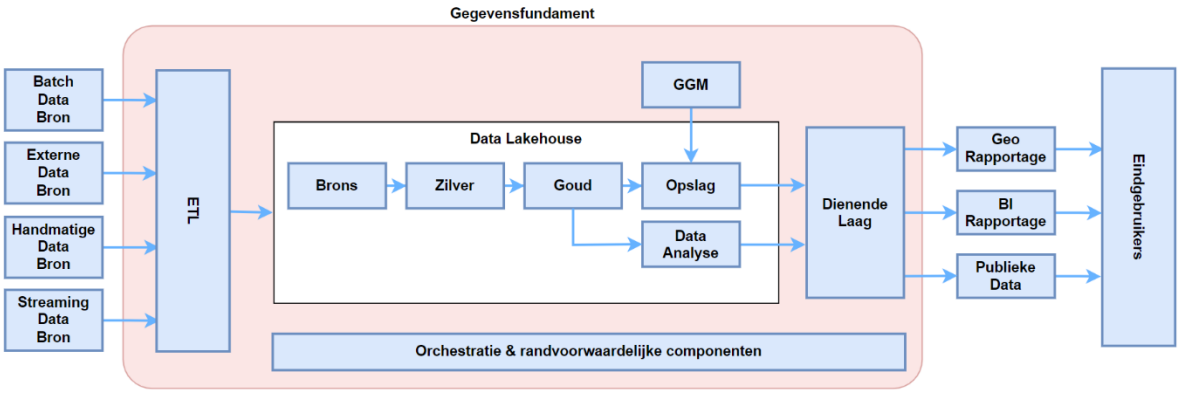
In bovenstaande figuur zijn de algemene componenten weergegeven waarbij een duidelijke generieke datastroom te herleiden is.

COMPONENT	TOELICHTING
Databronnen	Alle (externe) bronnen die data kunnen leveren aan het gegevensfundament.
Data Inladen	Ruwe data wordt ingeladen uit de bron (injectie), dit kan zowel een volledige bronontsluiting zijn of een selectievere ontsluiting van gegevens
Data Opslag	Opslag van (ruwe) data inclusief metadatering
Data Bewerken	Met één of meer bewerkingsslagen (hops) wordt de data vormgegeven zodat het doelmatig gebruikt kan worden
Data modelering	Component waarin de data op basis van een generiek model wordt ingedeeld

Dienende Laag	De faciliterende laag die de visualisatie laag bediend. In scope zit het 'bedienen' van de reeds in gebruik zijnde presentatie- en analysetools als PowerBI, ArcGIS en SPSS
Visualisatie laag	In deze laag bevinden zich componenten en applicaties die eindgebruikers in staat stellen gewenste informatieproducten te ontwikkelen (bijvoorbeeld: dashboard of kaartlaag).
Orkestratie & randvoorwaardelijke componenten	Ondersteunende laag waarin aspecten als regie, sturing, privacy, security, governance, etc. worden gefaciliteerd.

3.2 Doelarchitectuur

Op basis van de in 3.1 genoemde algemene componenten is de volgende doelarchitectuur opgesteld:



Er zijn een aantal uitgangspunten bepaald; het gegevensfundament:

- Ondersteunt verschillende bronnen (batch, streaming) op basis van verschillende stromen;
- Is gepositioneerd in de Azure Cloud;
- Volgt een Lakehouse architectuur;
- Laadt gegevens in op uniforme wijze met ETL;
- Voorziet in veilig gebruik van (gevoelige) data;
- Heeft een datamodel gebaseerd op het GGM;
- Faciliteert verschillende type eindgebruikers;
- Is in staat om op een veilige en uniforme manier data intern en extern te delen;
- Bestaat uit één omgeving voor de drie gemeenten;
- Voorziet in datascheiding waar nodig.

Ten aanzien van de oorspronkelijke doelarchitectuur uit 2021 zijn er een aantal wijzigingen, waarbij de voornaamste verschillen zijn samen te vatten in de volgende punten:

1. Het gegevensfundament is een volledige (public) cloudoplossing, en is niet op basis van een hybride Cloud omgeving.
2. Er wordt uitgegaan van een (medaillon)Lakehouse architectuur, in plaats van een traditionele Data Warehouse architectuur. Het medaillon principe maakt via een aantal 'hops' (van brons, naar zilver, naar goud) bewerkingsslagen mogelijk waarbij data opgeslagen, gevalideerd, en verrijkt wordt.

3. Pseudonimiseren van gegevens is geen generieke functionaliteit die wordt toegepast bij het inladen van gegevens; zie 3.2.1 voor toelichting.
4. Er wordt geen afzonderlijke API-gateway als component opgenomen in het gegevensfundament. Volgend aan de architectuur visie API-management wordt dit een generieke functionaliteit die ook ingezet gaat worden voor het gegevensfundament.

3.2.1 Pseudonimiseren

De startarchitectuur voorziet hierin door tijdens de gegevensbewerking (in een van de hops) daar waar nodig te pseudonimiseren. Dit biedt flexibiliteit en robuustheid zonder in te hoeven boeten op wensen en eisen vanuit een privacy perspectief.

In de bronzen laag wordt de benodigde ruwe, onbewerkte data uit de bronapplicatie opgeslagen. De toegang tot deze laag wordt tot een minimum beperkt. IN de hop van brons naar zilver wordt de data gepseudonimiseerd. Bij de hop van brons naar zilver is pseudonimiseren één van de datamutatieslagen. Wanneer er op termijn analyses uitgevoerd moeten worden of informatieproducten moeten worden geleverd waarbij het wel noodzakelijk is om herleidbare gegevens te kunnen gebruiken biedt deze structuur verschillende oplossingen om deze beschikbaar te stellen:

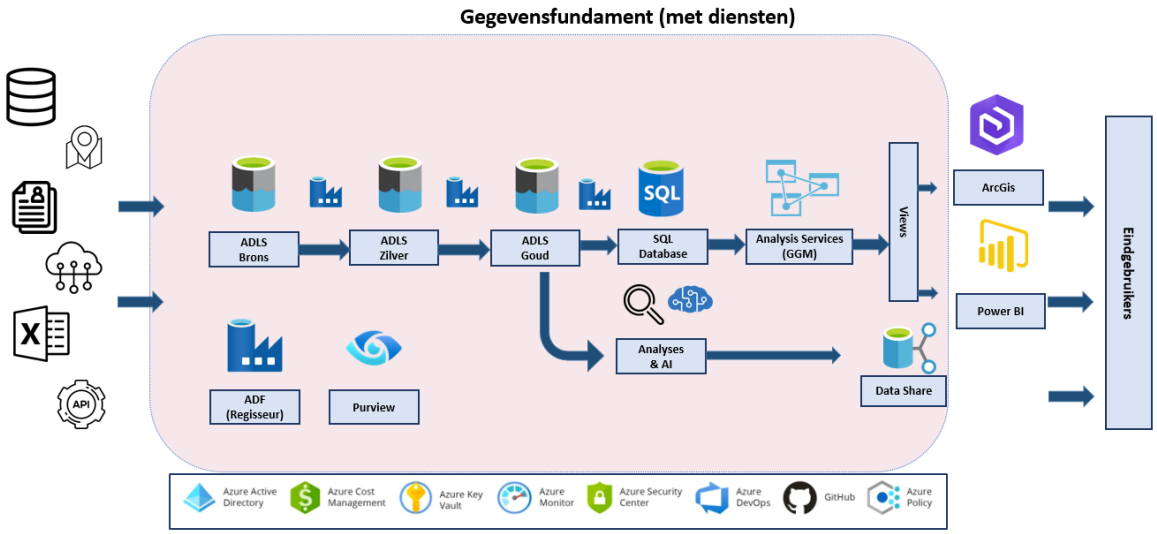
- Er kan binnen het gegevensfundament een aparte (afgescheiden) omgeving geplaatst worden waarbij de pseudonimiseringsslag niet wordt doorgevoerd. Op basis van autorisatie kan toegang tot deze omgeving worden verleend
- In de datamutatieslag wordt de gegevensset uitgebreid waardoor zowel de niet gepseudonimiseerde data als de gepseudonimiseerde data wordt opgeslagen. Op basis van autorisaties wordt bepaald welk type gegevens de gebruiker mag raadplegen.

De keuze voor een oplossingsrichting wordt onder andere bepaald door het structurele karakter van de onderliggende vraag, de gevoeligheid van de gegevens, en de technische impact. De keuze kan per onderliggende vraag verschillen.

Voor het gegevensfundament geldt in breedste zin het principe: ‘Data is gepseudonimiseerd, tenzij…’, waarbij de verwachting is dat in de startfase en eerste periode van gebruik geen uitzonderingen zullen plaatsvinden.

3.3 Startarchitectuur

Op basis van de vastgestelde componenten en de bouwblokken is het mogelijk om een oplossingsrichting met Azure diensten en componenten te beschrijven.



Deze startarchitectuur is gebaseerd op de nu bekende onderdelen en is niet limitatief (en kan dus ook niet als solution architectuur gezien worden). Het dient ter illustratie en verdieping, mede om als startpunt richting de aanbesteding gebruikt te kunnen worden. De genoemde diensten worden in dit voorstel niet op inhoud verder toegelicht.

Het portfolio van Azure is dermate divers dat de bovenliggende doelarchitectuur op verschillende manieren (en met verschillende diensten) kan worden ingevuld. Onderdeel van de aanbesteding is het opstellen van de definitieve solution architectuur.

In de weergegeven oplossingsrichting wordt gebruik gemaakt van een Azure Data Lake. Alternatieve oplossingsrichtingen zijn bijvoorbeeld op basis van Azure Synapse, Data Bricks, Snowflake, etc.

3.3.1 Microsoft Fabric

Voortbordurend op de afweging tussen PaaS en SaaS is de introductie en ontwikkeling van Microsoft Fabric een onderwerp van interesse. Fabric bevindt zich momenteel in een preview fase en past niet in het beoogde tijdspad. Op termijn kan een migratie richting Fabric een interessante keuze worden. De keuze voor een Lakehouse architectuur is hierbij een voorbereidende stap.

3.4 Randvoorwaarden

De positionering van het gegevensfundament in een Azure omgeving stelt een aantal randvoorwaarden.

3.4.1 Management Subscription/ Landing Zone

Op dit moment is er geen ingerichte Azure omgeving die ingezet kan worden voor de ontwikkeling van het gegevensfundament. De afhankelijkheid en impact op doorlooptijden is in een apart memo verder toegelicht.

3.4.2 API Management

Parallel aan het opstellen van het gegevensfundament is er een visie opgesteld ten aanzien van generiek API-management. In deze visie wordt een centraal integratie platform in het landschap gepositioneerd dat voorziet in interne en externe ontsluitingen.

Op termijn is de afhankelijkheid tussen API-management en het gegevensfundament duidelijk, maar gezien de discrepantie tussen de doorlooptijden is het mogelijk dat niet op voorhand alle koppelingen via een centraal integratie platform lopen. API-management is in die zin rand voorwaardelijk, maar niet per definitie blokkerend voor het gegevensfundament.

In het kader van aanbesteding (en zicht houden op eventuele architectuurschuld) is het van belang de afhankelijkheid kenbaar te maken.

Bijlage 1: Beheertaken op hoofdlijnen

Rol	Taken
Information Security Officer	▪ Uitvoeren DPIA op applicatieniveau
Privacy Officer	▪ Uitvoeren DPIA bij ontsluiten nieuwe bronnen ▪ Coördinatie uitvoering DPIA
Technisch beheer	▪ Monitoring systeemcapaciteit en prestaties en het treffen van maatregelen om dit te optimaliseren ▪ Monitoring (informatie)veiligheid en het treffen van maatregelen om dit te verbeteren ▪ Back ups en updates uitvoeren ▪ Beheer firewalls
Functioneel (applicatie) beheer	▪ Gebruikersondersteuning ▪ Beheer toegang (autorisaties) ▪ Beheren van de inrichting (functionaliteit) van de applicatie ▪ Uitvoeren (gebruikers)testen
Engineer gegevensfundament	▪ Beheer van het datamodel ▪ Meta data management ▪ Vernietiging van data