Introductie DIM architectuur

Project DataFabriek
29 Augustus 2023

Agenda

Aanleiding project DataFabriek

Het Data Integratie Magazijn

Bronzone

Integratie Zone

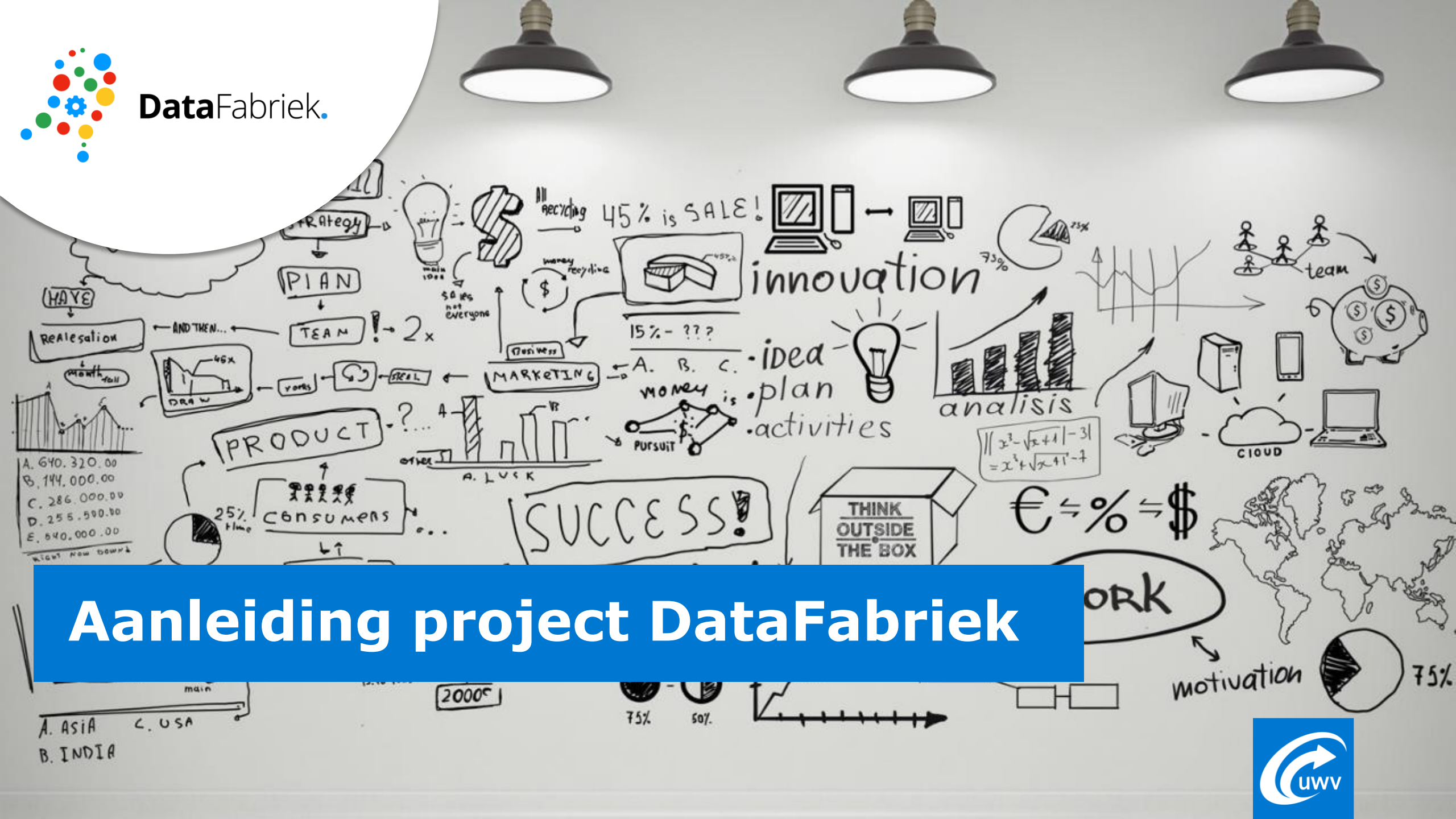
Bedrijfszone

Metadata en lineage

Privacy maatregelen

(nog meer) Vragen

DataFabriek.

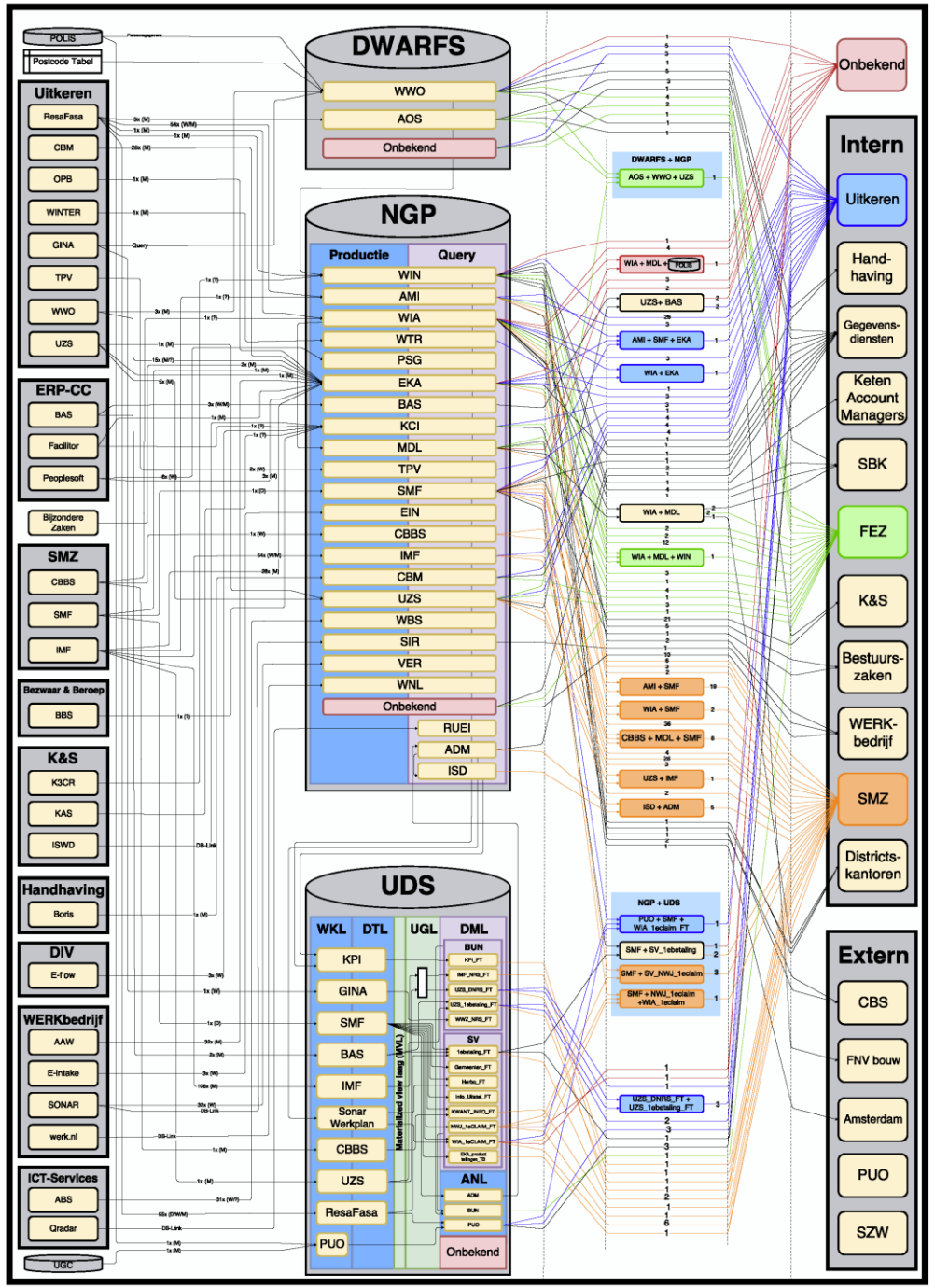


Aanleiding project DataFabriek

uwv

Oeps..

Datawarehouse UWV - 1 oktober 2017



Drijfveren project uit de PSA

Hoge prioriteit

Garanderen stabiliteit & continuïteit van de bestaande dienstverlening

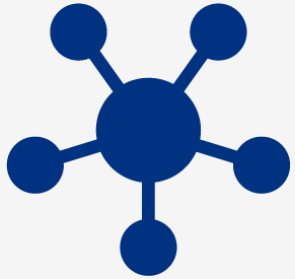
Verkleinen time-to-market informatieverzoeken

Verbeteren gegevensbescherming

Later in het project (maar wel noodzakelijk)

Meten / verbeteren datakwaliteit

Ondersteunen nieuwe typen (“self service”) gegevensgebruik



Ondersteunen nieuwe typen data (b.v. ongestructureerd)

DataFabriek.

IST - Continuïteits- en compliance risico's

Drie centrale datawarehouses

- DWH 1.0 (DWARFS)
  Gebaseerd op COBOL maatwerksoftware
- DWH 2.0 (NGP)
  Gebaseerd op maatwerk PL/SQL-generatoren
- DWH 3.0 (UDS)
  Gebaseerd op Oracle Warehouse Builder OWB

Alle drie met continuïteitsrisico's

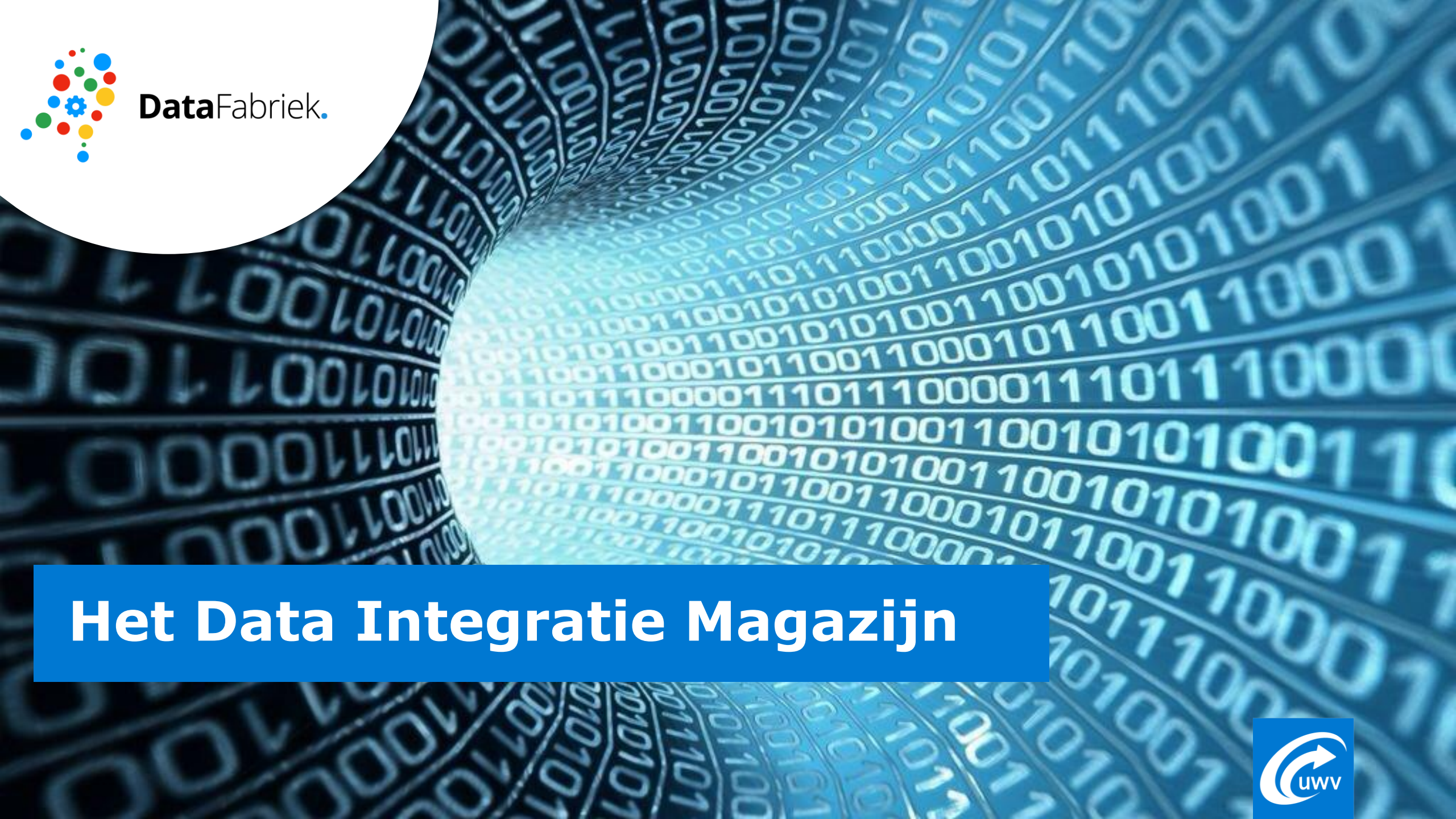
- DWH 1.0 / 2.0
  Kennis van de onderliggende tooling is nog maar beperkt aanwezig
- DWH 3.0
  OWB sinds juli 2018 niet meer in premium support.
  De mogelijkheid voor extended support eindigt in juli 2021
  (N.B. termijn nu verlengd tot eind 2022)

Daarnaast compliance-risico's

- Pseudonimisering / anonimisering wordt door huidige toolsets niet (of slechts beperkt) ondersteund.
- Traceerbaarheid ("horizontale lineage") wordt door huidige toolsets niet ondersteund

DataFabriek.

Het Data Integratie Magazijn



Uitgangspunten oplossingsarchitectuur

Risicominimalisatie

- Bewezen concepten en producten
  - Geleerde lessen uit andere organisaties
  - Geen “cutting edge” als fundament
- KISS (Keep it simple, stupid)
  - Van-de-plank, tenzij
  - Heldere regels en verantwoordelijkheden
- Stapsgewijze migratie
  - Per informatieproduct of “use case”
  - Maar wel binnen één doelarchitectuur
  - Wegwerken continuïteitsrisico’s met stip op 1
  - Zonder de langere termijn uit het oog te verliezen

Toekomstvastheid

- Wendbare oplossing:
  - Maximale ontkoppeling van lagen & oplossingscomponenten
  - Ontwikkel- en beheerstandaards (incl. patronen en hergebruik)
- Klaar voor nieuwe technologie:
  - Big data, data virtualisatie
  - Cloud
- AVG-compliant
  - Privacy by design/default
- Schaalbaar
  - Parallelle verwerking, delta-processing

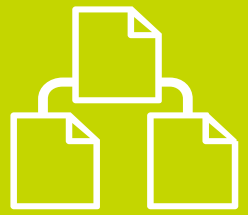
Producten en diensten “Datafabriek”

Gegevens-leveringen

- Datamarts
- Gegevensvensters
- Gegevensbestanden

Beheer & inrichting

- Inrichten gebruiksomgeving
- Beheer op eindgebruikers-middelen
- Beheer op levering



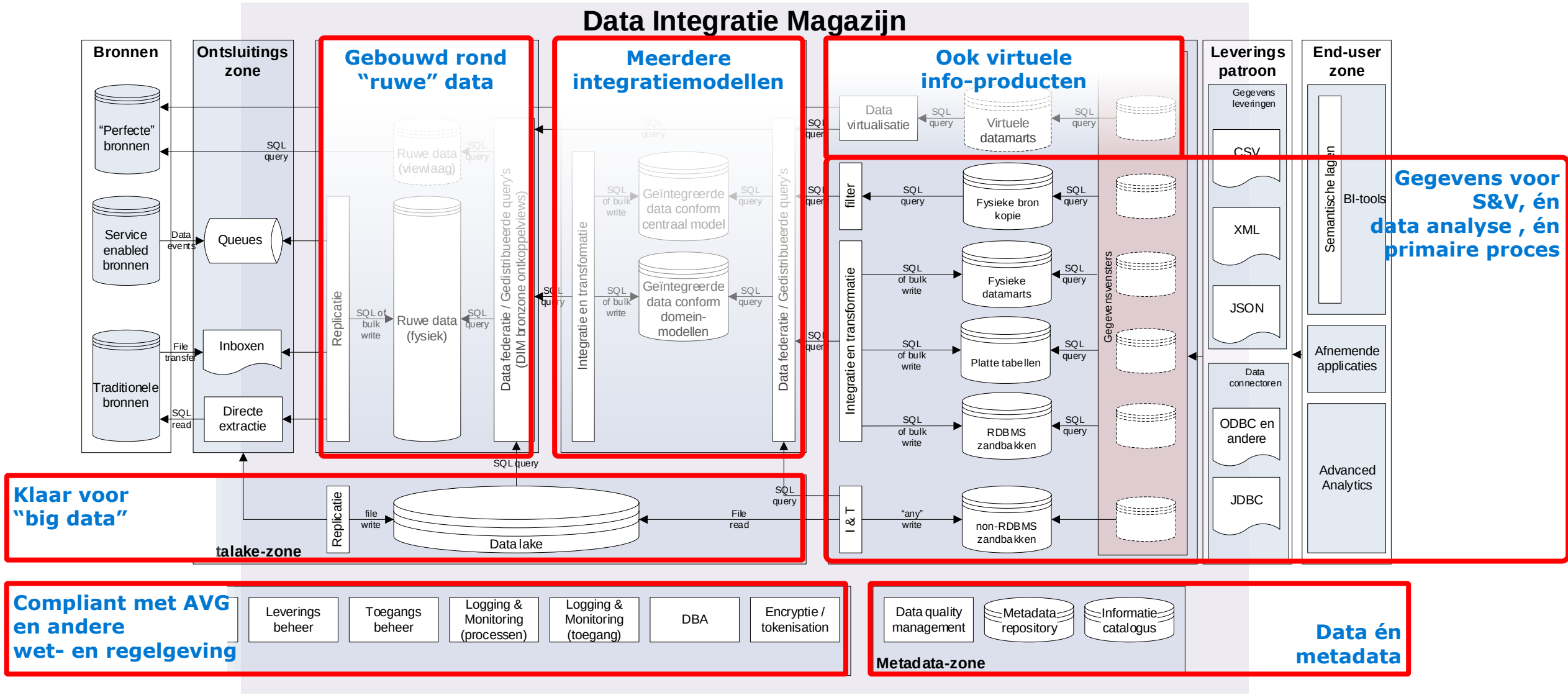
Compensatie

- Compensatie historie
- Compensatie beschikbaarheid
- Compensatie snelheid

Expertise-diensten

- Inzicht kwaliteit
- Inzicht meta-gegevens
- Rapportagedienst
- Expertisedienst

Het DIM is een "state of the art" Enterprise Data Warehouse




DataFabriek.
Bronzone
uwv

Bronzone

Het doel van de bronzone is om een historische representatie van de administratieve werkelijkheid van een bron beschikbaar te maken.

- Het DIM dient o.a. om de beperkingen van de bronsystemen m.b.t. historisch besef en de beschikbaarheid van gegevens voor het gebruik in data integratie/analyse (DIA) te “compenseren”.

- Hiervoor worden gegevens uit de bronsystemen in kopie (en zonder verdere bewerkingen) in het DIM opgeslagen. Als een bronsysteem geen gegevenshistorie bijhoudt (m.a.w. bij wijziging van gegevens de oude waarden overschrijft), dan wordt deze gegevenshistorie in het DIM alsnog opgebouwd.

- De gegevens in het DIM moeten, voor DIA-gebruik, kunnen worden beschouwd als “identiek” aan die in bron. Hieruit volgt een aantal eisen m.b.t. de gegevensleveringen aan het DIM.

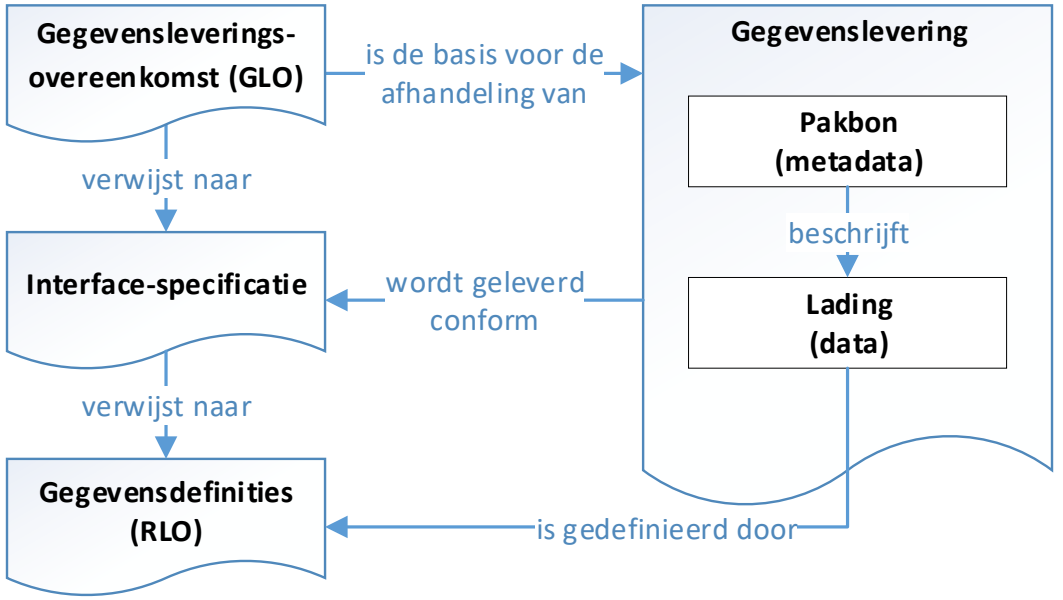
- Daarnaast stellen ook wet- en regelgeving (m.n. de AVG) en het eigen UWV-beleid t.a.v. gegevensbeheer eisen aan replicatie naar en opslag in het DIM. (o.a. DLM, maskering en beveiliging)

Zes hoofdeisen voor de gegevensleveringen aan het DIM

Voor de aan het DIM geleverde gegevens	Voor de wijze van levering van die gegevens
G1 - De gegevens zijn volledig	L1 - De levering is compliant
G2 - De gegevens zijn gedefinieerd	L2 - De levering is robuust
G3 - De gegevens zijn herleidbaar	L3 - De levering is toekomstvast

Aandachtsgebieden voor de interface standaarden

- De interface standaarden dekken functionele, technische, documentaire en procedurele eisen aan gegevensleveringen van bron naar DIM.
- Ze betreffen:
  - De via de interface te leveren gegevens (de “interface-lading”)
  - De via de interface te leveren metadata (de “interface-pakbon”)
  - De fysieke implementatie van de interface (bij de bron)
  - De diepgang waarmee en wijze waarop de functionaliteit van de interface vastgelegd.
  - De diepgang waarmee en de wijze waarop definities van de via de interface geleverde gegevens worden vastgelegd.
  - De noodzakelijke procesafspraken rondom de interface en hoe deze in gegevens inwinning afspraak (GIA’s was eerder de GLO) worden vastgelegd.



Beperkte eisen t.a.v. formaat van gegevens en levering

Bronzone

Gelaagdheid in de verwerking naar de bronzone

- Eerst de levering naar de staging
- Kwaliteitscheck van de levering zodat we weten dat we deze correct kunnen verwerken naar de bronzone
- Verwerking van de staging tabellen naar de bronzone tabellen (maskering vindt in deze stap plaats)

Harnassen (werkend op stuur metadata gebaseerd op de informatie in de RLO)

- Platte bestanden met pakbon
- Oracle datapump met par en log bestanden
- Database link
- Toekomstig: berichten
- Staging naar de bronzone

Ontkoppelviews

- Gemaskeerd
- Ongemaskeerd
- Niet identificerend

Maskering

Maskering

- Maskeringsmatrix met klassen
- Identificerende gegevens
- Voor koppelbare gegevens kan uniformering noodzakelijk zijn

Maximale pragmatische maskering

We moeten voorkomen dat mensen om hun werk te kunnen doen als enige optie hebben om de niet gemaskeerde gegevens uit het DIM te gebruiken. In de meeste gevallen zouden mensen met de gemaskeerde versie moeten toekunnen.

Bijv. geboortedatum

Maskeringsmethoden

In het DIM worden, bij opslag, de volgende maskeringsmethoden toegepast:

Vervangen (koppelbaar)	Een gegeven ongemaskeerde waarde voor een gegevenselement krijgt altijd dezelfde unieke 'gemaskeerde' waarde, voor alle velden waarin dit gegevenselement voorkomt. Hierdoor blijft het koppelen van gegevens op basis van deze velden mogelijk. N.B. Koppelen via "fuzzy matching" ("Jansen" ongeveer gelijk aan "Janssen") is na maskeren niet meer mogelijk.
Vervangen (willekeurig)	In dit geval kan de 'gemaskeerde' waarde van een bepaalde waarde elke keer anders zijn. Koppelen op basis van op deze wijze gemaskeerde velden is dus niet mogelijk.
Verbergen	De waarde van het veld wordt, geheel of gedeeltelijk, vervangen door een standaard-karakter of -cijfer. *Voorbeeld: Van een IBAN blijven de eerste 8 karakters ongewijzigd, en wordt de rest van het rekeningnummer overschreven met "X".*
Volledig verbergen	Een extreem geval van **verbergen**. Hierbij wordt het gehele veld gevuld met "X".
Volledig leeglaten	Een extreem geval van **verbergen**. Hierbij wordt het gehele veld in de gemaskeerde gegevens leeg gelaten.
Classificeren	De waarde van het veld wordt op basis van een 'algoritme' vervangen door een waarde uit een beperktere reeks waarden. *Voorbeeld: Postcode in de ongemaskeerde data vervangen door Postcodegebied in de gemaskeerde data.*
Indicator gevuld	Een extreem geval van **classificeren**. Hierbij wordt het gemaskeerde veld tot alleen de waarden "gevuld" of "leeg" beperkt.

Data Lifecycle Management

Het DIM volgt de DLM van de bron met betrekking tot de niet gemaskeerde variant van identificerende gegevens. (m.n. AVG compliance en Archiefwet)

Eerst een soft delete en na propagatie naar de integratie en bedrijfszone en het verlopen van de grace periode om herstel mogelijk te maken wordt het een hard delete.

Voor de overige gegevens geldt een bewaartermijn van 20 + 5 jaar (m.n. door de Archiefwet)

Lineage

Omdat de ontkoppelview gebaseerd op de bronzone van het DIM een historische representatie van de administratieve werkelijkheid van de bron weergeven is elk attribuut volledig gerelateerd aan hoe deze is geleverd door de bron.

Dus de ontkoppelviews vormen de basis van de voor de afnemers beschikbare lineage (business lineage)

Technisch is er nog een andere lineage beschikbaar welke minder duidelijk is door de metadata gedreven harnassen.


DataFabriek.
Integratie zone
uwv

Integratie Zone

Informatie gebied vs informatie product

- Een informatie gebied is de groepering van gegevens die een bepaalde functioneel vraag beantwoord.
- Een informatie product is de beschikbaar stelling van (een deel van) het informatie gebied. Hierbij kunnen er bijv. nog additionele filteringen (o.a. VIPs, additionele maskering/filtering) plaats vinden.

Wat doen we in de integratie zone?

- Integratie van gegevens uit tabellen binnen een bron
- Integratie van gegevens uit tabellen over bronnen heen
- Afleidingen van gegevens (vaak door middel van business rules)
- Toepassen DLM uit de bronzone
- Custom ETL (om lineage helder te kunnen laten zien)


DataFabriek.
Bedrijfszone
uwv

Bedrijfszone

Wat doen we in de bedrijfszone?

- Gegevens klaar zetten in de optimale vorm om bruikbaar te zijn voor de afnemers
  - Datamarts
  - Platte bestanden
  - …
- Toepassen DLM uit de bronzone
- Custom ETL (om lineage helder te kunnen laten zien)

Datamart/platte tabellen/gegevensvensters

Een gegevensvenster vormt de toegang tot gegevens in het DIM

- Ontkoppelviews van de bron (ruwe data. Gemaskeerde of niet gemaskeerde identificerende gegevens, niet identificerend gegevens)
- Integratie zone van een informatie gebied
- Bedrijfszone (datamart / platte tabellen)

Additionele maatregelen om gegevens af te schermen

- Toegang gaat op basis van doelbinding en rechtsgrond.
- Een persoon kan alleen toegang krijgen tot de gemaskeerde of niet gemaskeerde variant van identificerende gegevens. Nooit tot beide. Als wel leidt dit tot een automatische melding van een gegevenslek omdat de beveiliging van gegevens in het DIM gecompromitteerd is.
- VIP filtering vindt plaats indien een persoon geen VIP gegevens mag zien (geldt alleen voor niet gemaskeerd)
- Eigen medewerkers
- Alleen gegevens van het eigen kantoor
- …


DataFabriek.
Metadata
uwv

Metadata en lineage

Metadata en lineage van informatie gebieden wordt beschikbaar gesteld aan de afnemers

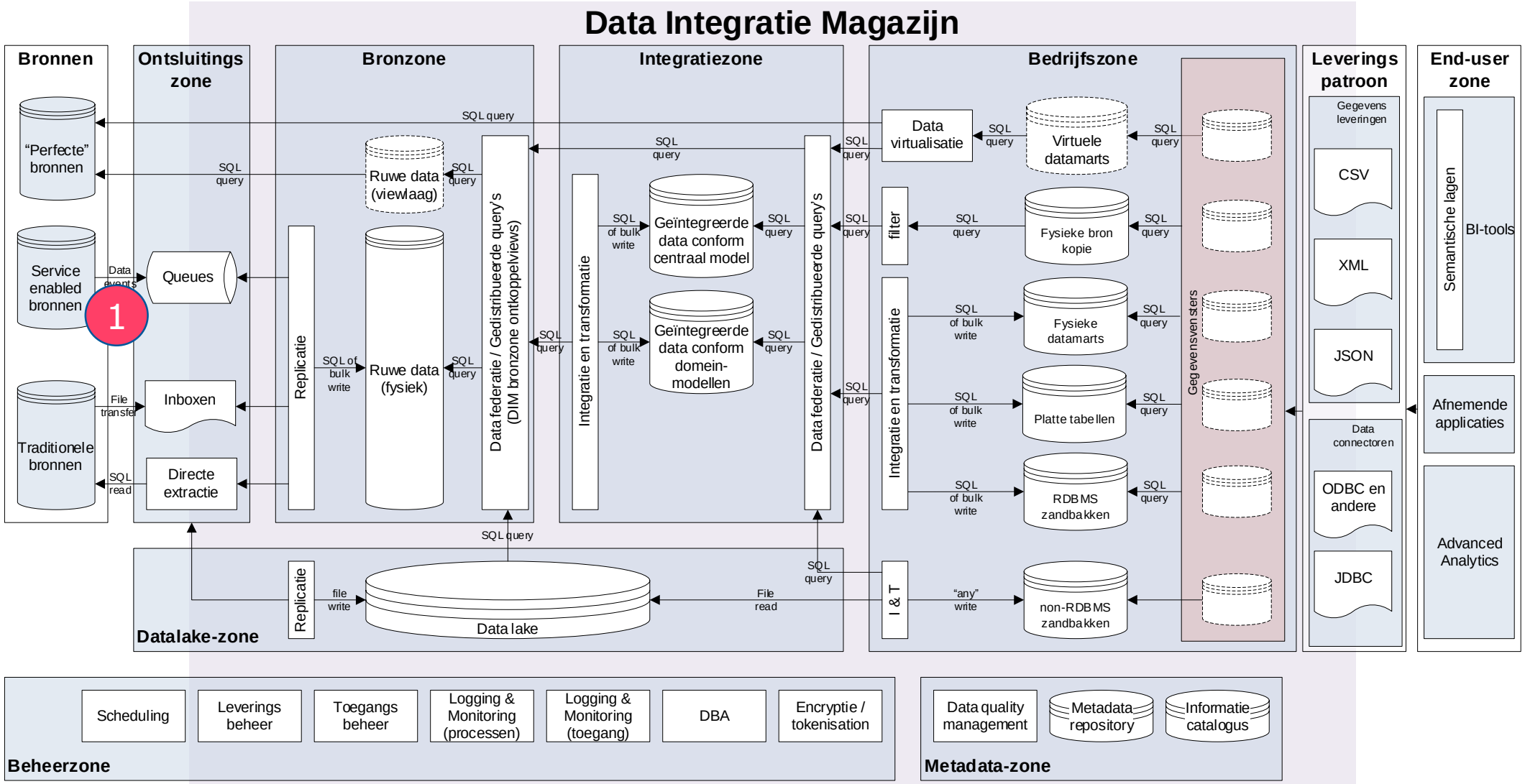
Dit is iets dat nieuw is ten opzichte van de bestaande DWH’s

De metadata geeft de afnemer de functionele betekenis van een gegeven

De lineage laat de afnemer zien hoe het gegeven tot stand is gekomen en op basis van welke gegeven(s) uit welke bron dit is.


DataFabriek.
Privacy maatregelen
uwv

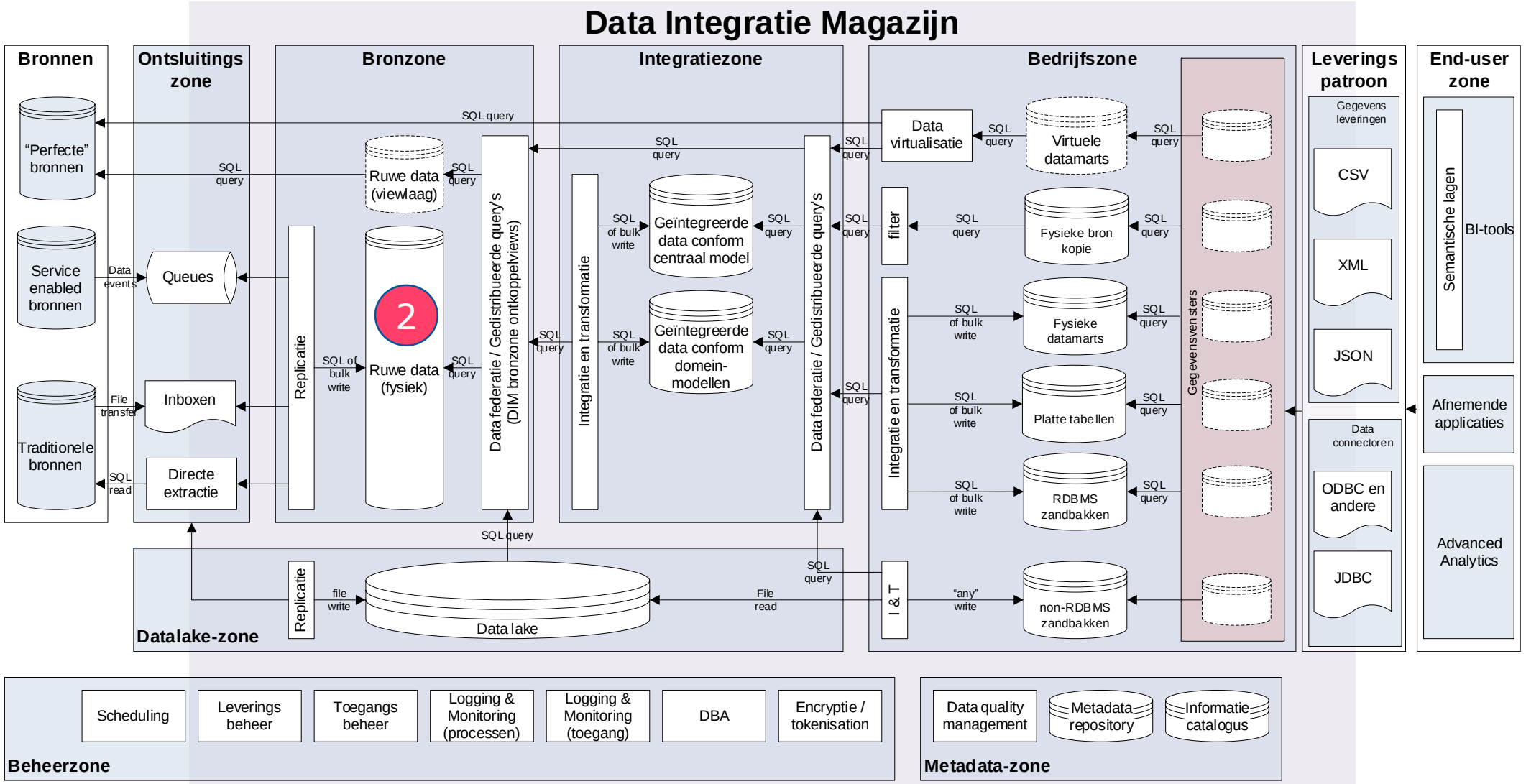
DIM privacy maatregelen



Maatregelen bij de bron gegevenslevering

- Niet leveren indien ook werkelijk niet noodzakelijk (Filteren bij de bron)
- Gebruik beperkingen opnemen in de GIA
  - Wat mag wel
  - Wat mag niet
  - Wat mag eventueel met toestemming
- De bron blijft eigenaar van de geleverde gegevens

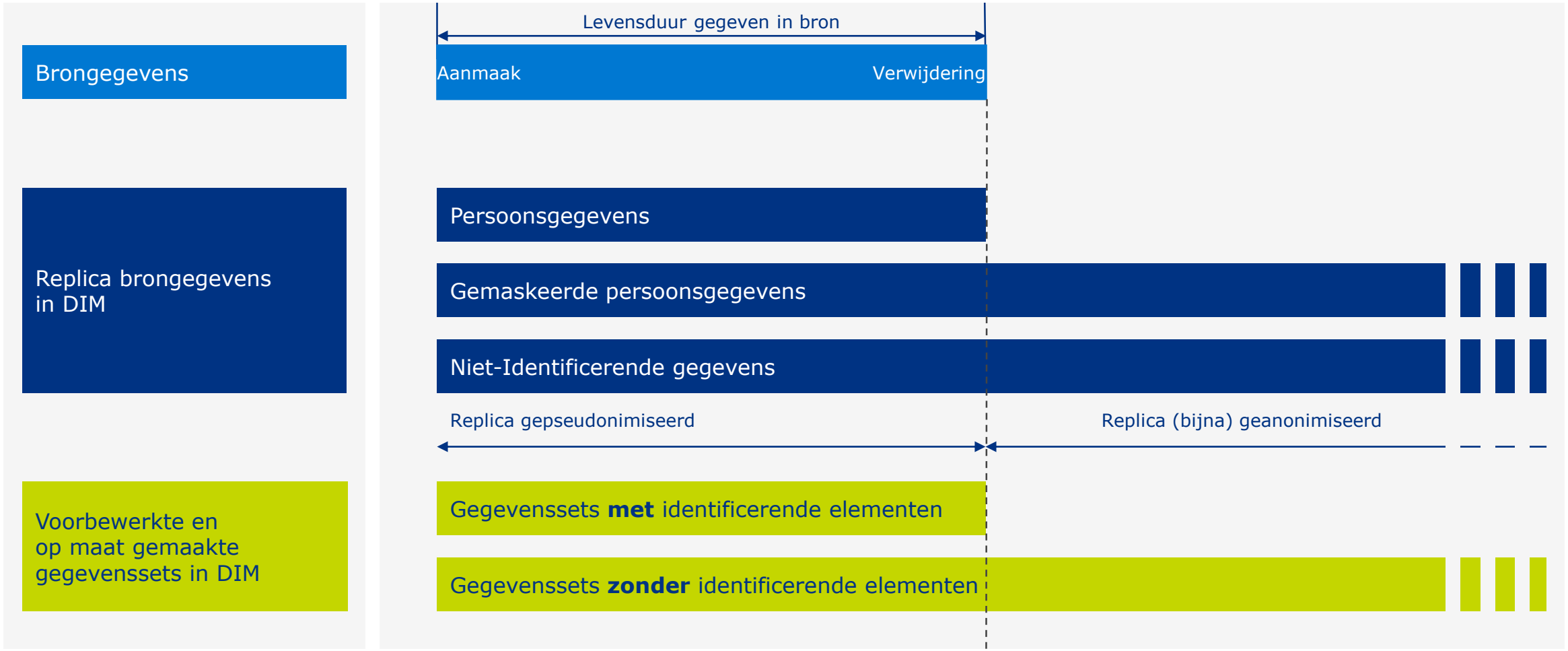
DIM privacy maatregelen



Maatregelen in de bronzone

- Identificerende gegevens worden gemaskeerd en niet gemaskeerd opgeslagen
- Gemaskeerde en niet gemaskeerde gegevens hebben een gescheiden DLM
  - Gemaskeerde gegevens worden typisch na 20 + 5 jaar verwijderd
  - Niet gemaskeerde gegevens worden verwijderd doordat we de DLM van het bronsysteem volgen

AVG & Data Lifecycle Management

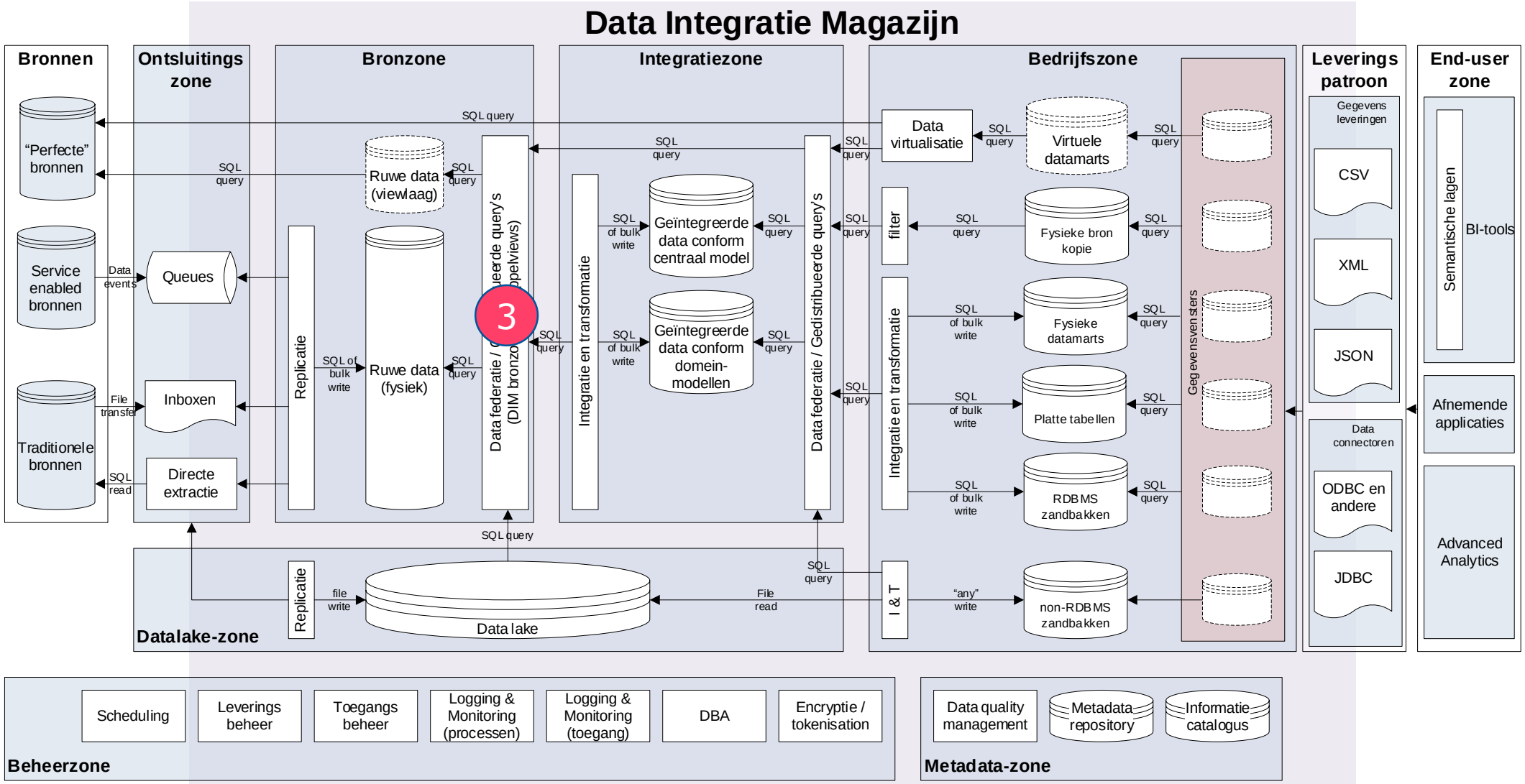


Maskeringsmethoden

In het DIM worden, bij opslag, de volgende maskeringsmethoden toegepast:

Vervangen (koppelbaar)	Een gegeven ongemaskeerde waarde voor een gegevenselement krijgt altijd dezelfde unieke ‘gemaskeerde’ waarde, voor alle velden waarin dit gegevenselement voorkomt. Hierdoor blijft het koppelen van gegevens op basis van deze velden mogelijk. N.B. Koppelen via “fuzzy matching” (“Jansen” ongeveer gelijk aan “Janssen”) is na maskeren niet meer mogelijk.
Vervangen (willekeurig)	In dit geval kan de ‘gemaskeerde’ waarde van een bepaalde waarde elke keer anders zijn. Koppelen op basis van op deze wijze gemaskeerde velden is dus niet mogelijk.
Verbergen	De waarde van het veld wordt, geheel of gedeeltelijk, vervangen door een standaard-karakter of -cijfer. *Voorbeeld: Van een IBAN blijven de eerste 8 karakters ongewijzigd, en wordt de rest van het rekeningnummer overschreven met “X”.*
Volledig verbergen	Een extreem geval van **verbergen**. Hierbij wordt het gehele veld gevuld met “X”.
Volledig leeglaten	Een extreem geval van **verbergen**. Hierbij wordt het gehele veld in de gemaskeerde gegevens leeg gelaten.
Classificeren	De waarde van het veld wordt op basis van een ‘algoritme’ vervangen door een waarde uit een beperktere reeks waarden. *Voorbeeld: Postcode in de ongemaskeerde data vervangen door Postcodegebied in de gemaskeerde data.*
Indicator gevuld	Een extreem geval van **classificeren**. Hierbij wordt het gemaskeerde veld tot alleen de waarden “gevuld” of “leeg” beperkt.

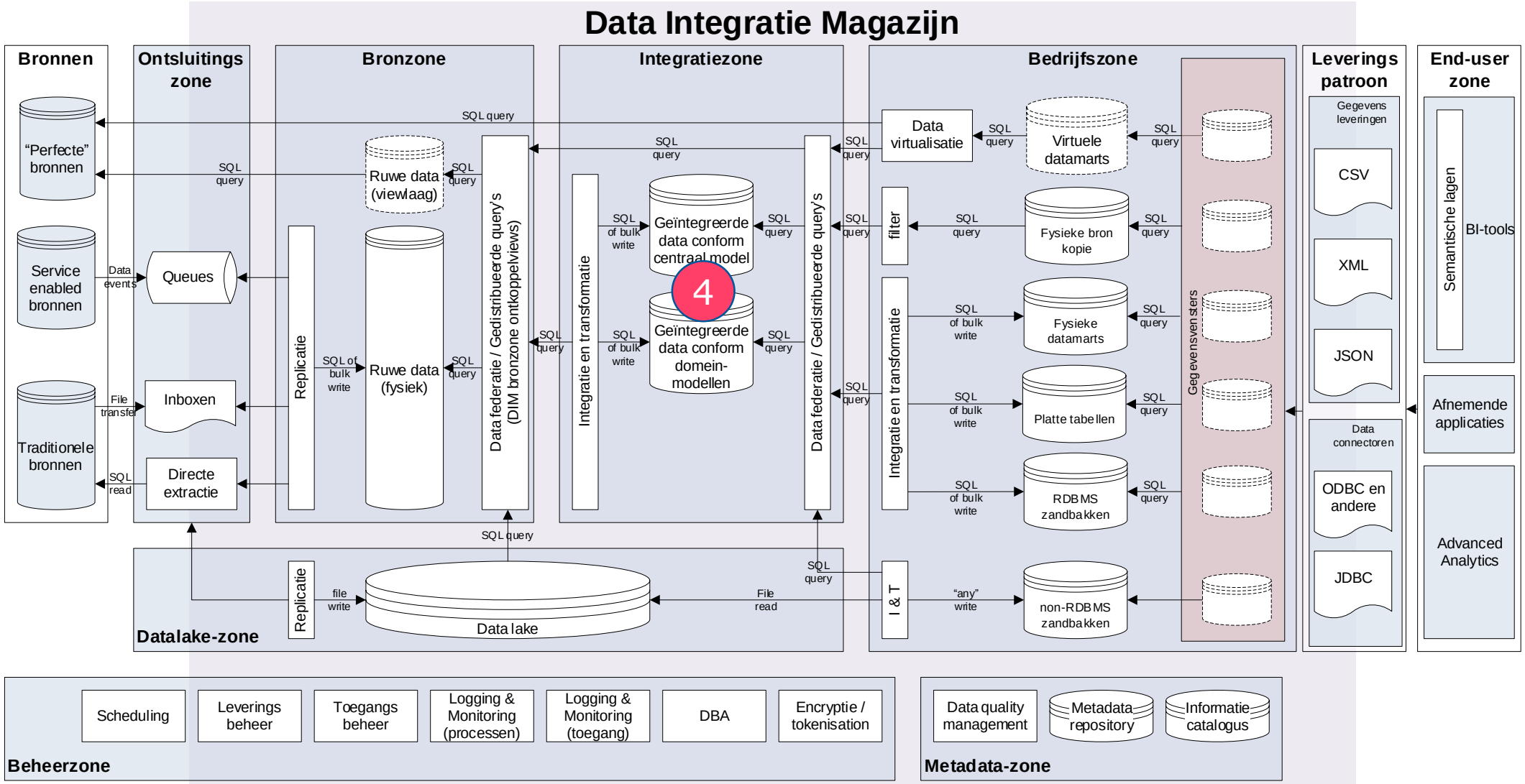
DIM privacy maatregelen



Maatregelen voor de ontkoppelviews

- Ontkoppelviews zijn gemaskeerd en niet gemaskeerd beschikbaar binnen het DIM
- Qua structuur zijn de gemaskeerde en niet gemaskeerde ontkoppelviews identiek
- Hierdoor kan voor de informatie gebieden van het DIM dezelfde ETL voor zowel de gemaskeerde als de niet gemaskeerde variant van een informatiegebied gebruikt worden.

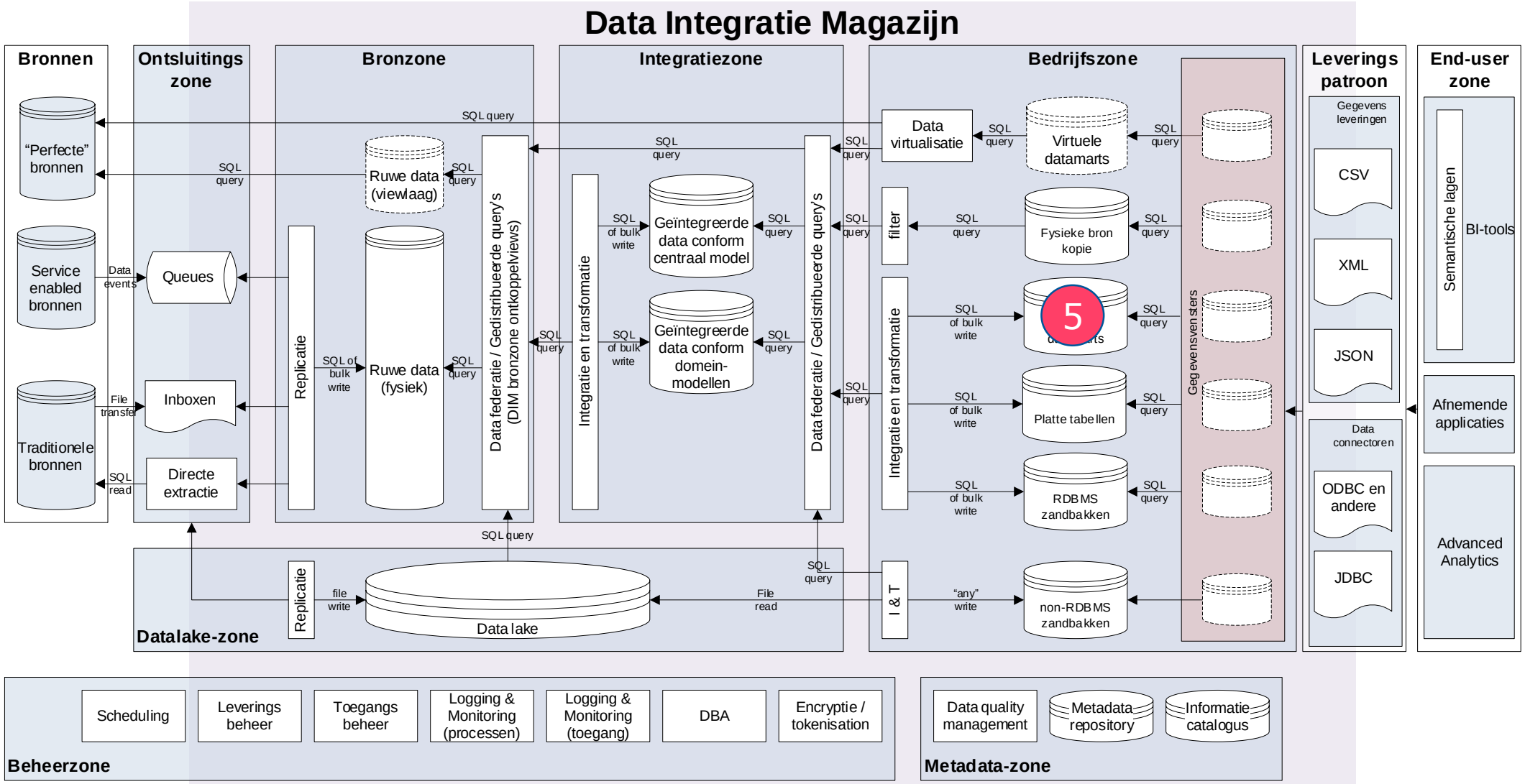
DIM privacy maatregelen



Maatregelen in de integratiezone

- Neem alleen mee wat benodigd is qua gegevens voor een informatie gebied
  - Standaard dus geen VK3 of andere gegevens waarvoor expliciete toestemming nodig is van de gegevenseigenaar (Zie RLO en GIA). Dit mag alleen als er een zwaarwegend belang is en toestemming van de gegevenseigenaar.
- Een informatie gebied wordt alleen in de benodigde varianten aangemaakt
  - Gemaskeerd en/of niet gemaskeerd (dus alleen wat nodig is)

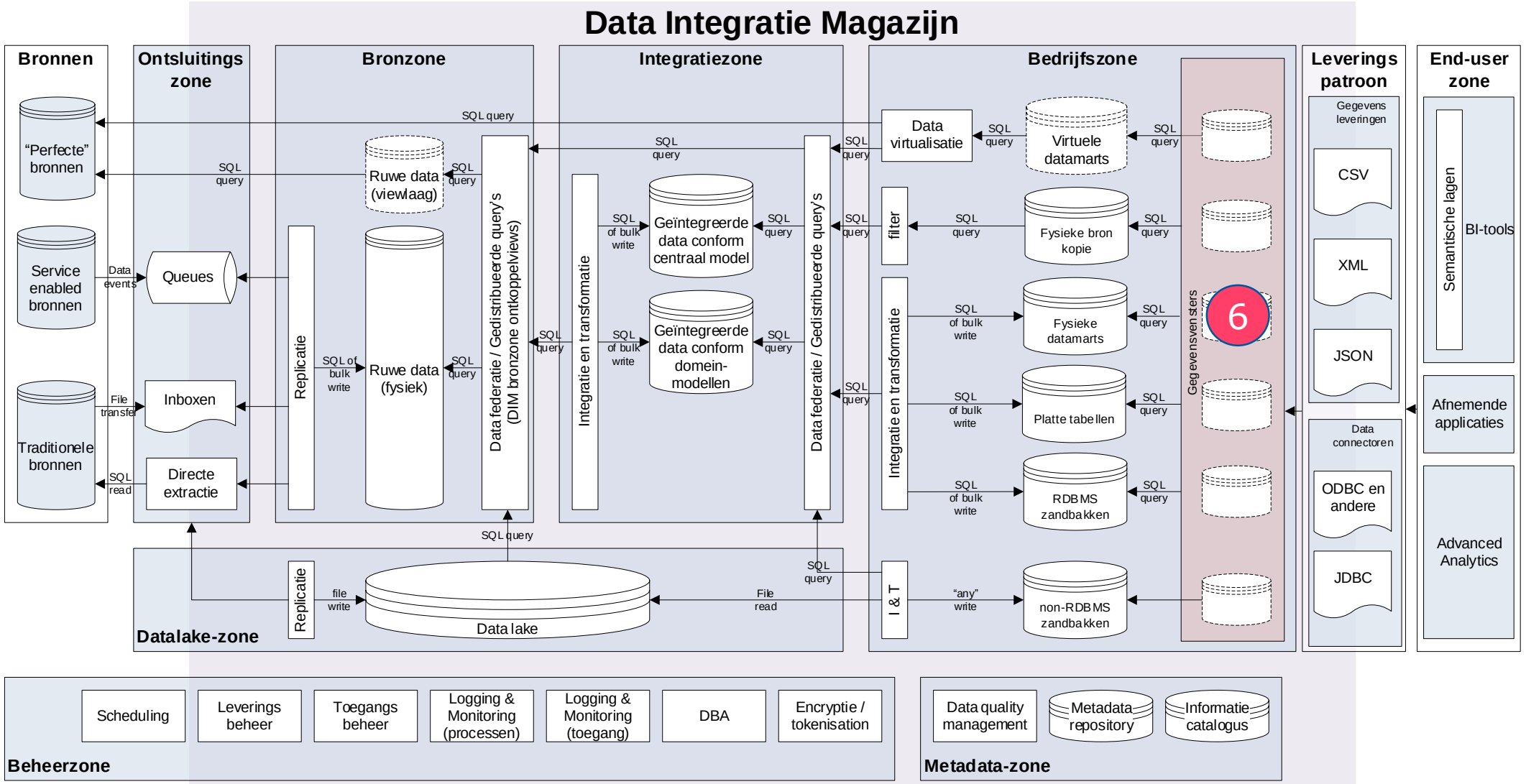
DIM privacy maatregelen



Maatregelen in de bedrijfszone

- Neem alleen mee wat benodigd is qua gegevens voor een informatie gebied
  - Standaard dus geen VK3 of andere gegevens waarvoor expliciete toestemming nodig is van de gegevenseigenaar (Zie RLO en GIA). Dit mag alleen als er een zwaarwegend belang is en toestemming van de gegevenseigenaar.
- Een informatie gebied wordt alleen in de benodigde varianten aangemaakt
  - Gemaskeerd en/of niet gemaskeerd (dus alleen wat nodig is)
- Indien mogelijk gegevens oprollen / verder anonimiseren

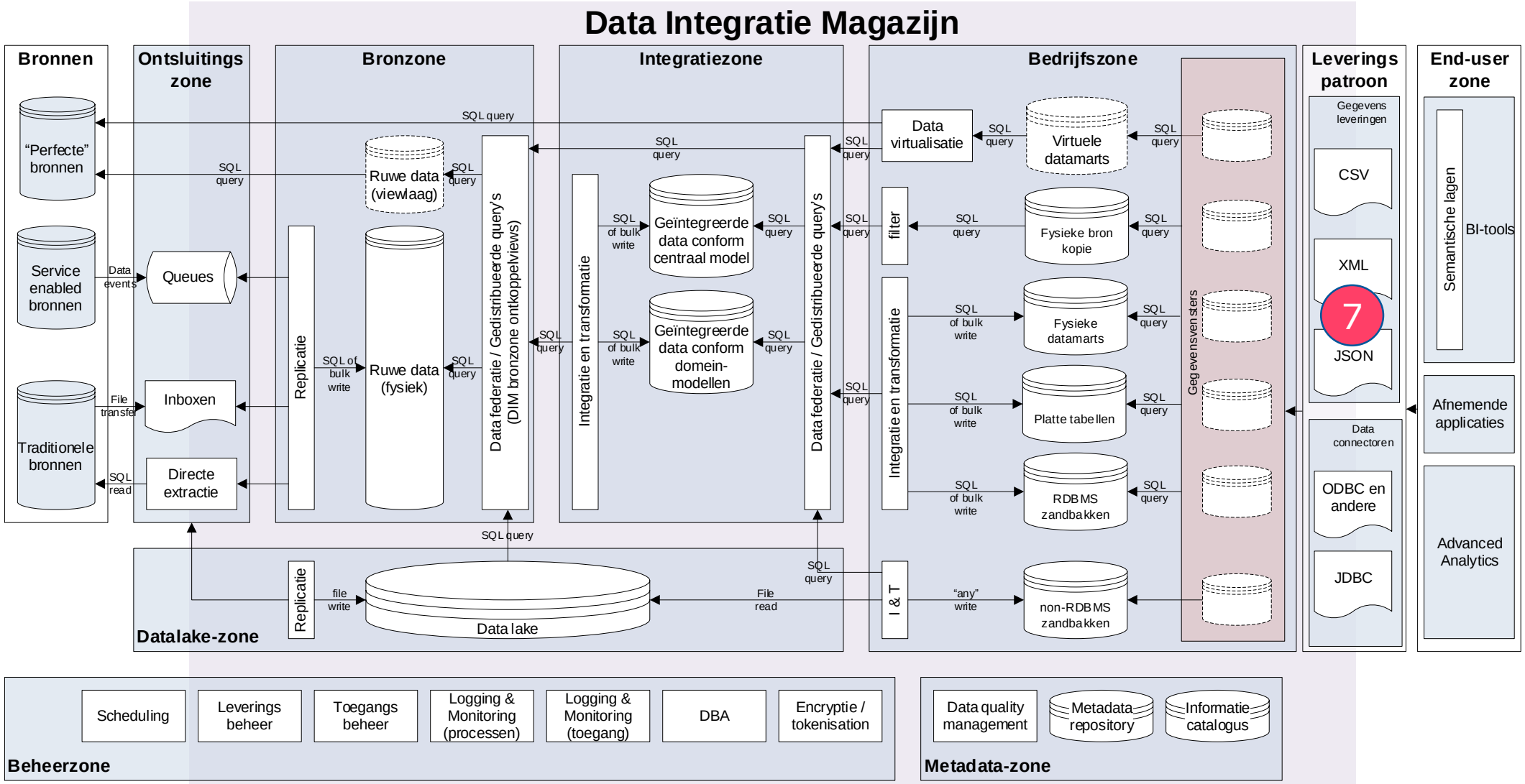
DIM privacy maatregelen



Maatregelen in de gegevensvensters

- Een gegevensvenster is de enige vorm van toegang tot gegevens in het DIM
- Een gegevensvenster is per definitie specifiek voor een rol van een afnemer waarbij gegevensminimalisatie wordt toegepast. Ofwel de afnemer krijgt niet meer gegevens te zien dan dat benodigd is voor het uitvoeren van zijn taken en verantwoordelijkheden binnen UWV.
- Gegevensminimalisatie conform afspraken in de GLA
  - Filteren van gegevens
  - Anonimiseren van gegevens
- Niet gelijktijdig toegang tot gemaskeerd en niet gemaskeerde gegevens om een datalek te voorkomen.
- Gemaskeerde gegevens (in het DIM niet geanonimiseerd) hebben nog steeds doelbinding, proportionaliteit en subsidiariteit noodzakelijk.

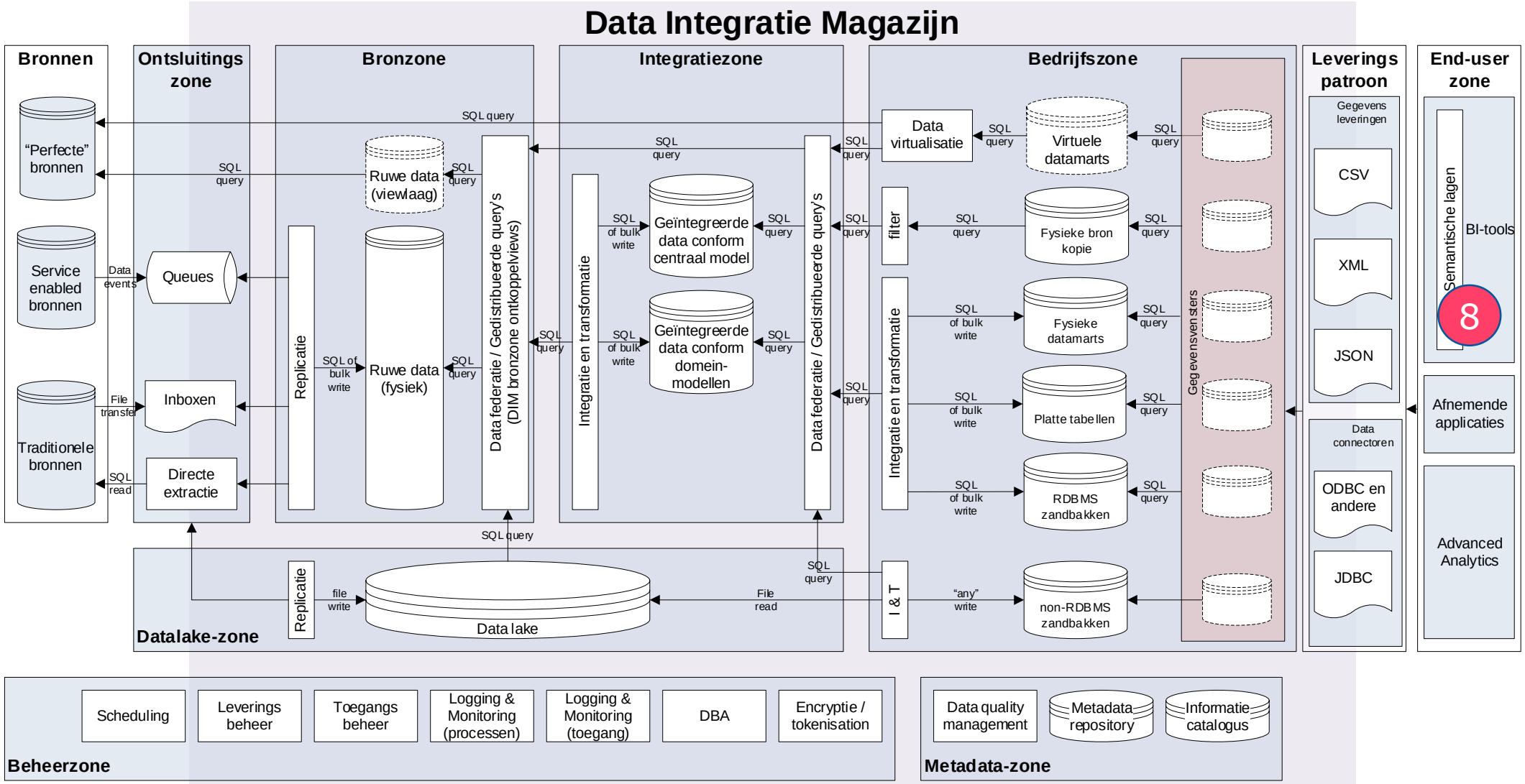
DIM privacy maatregelen



Maatregelen bij leveren / beschikbaar stellen

- Doelbinding proportionaliteit subsidiariteit / zwaarwegend belang (VK3) controle
- Gebruik beperkingen GLA
- Op gebruiker gebaseerde authenticatie / autorisatie

DIM privacy maatregelen



Maatregelen bij de afnemer (1)

- Gebruik beperkingen in de GLA

- Verantwoordelijk voor gegevensvernietiging van bestanden en rapportages conform UWV standaarden en richtlijnen

- In principe any BI en Analytics tooling kan gebruikt worden mits deze voldoen aan de UWV architectuur standaarden en richtlijnen voor het gebruik van gegevens in het DIM

Maatregelen bij de afnemer (2)

- Additionele privacy maatregelen (goedgekeurde GEB + bijbehorende inrichting indien het een niet bij Gegevensdiensten beheerde BI / Analytics tool betreft)
  - Op gebruiker gebaseerde authenticatie en autorisatie voor het ophalen van gegevens uit het DIM (default, dus zonder tussen opslag van gegevens)
  - Indien er een NON-GA gebruikt wordt mag dit alleen of gemaskeerde of niet gemaskeerde gegevens betreffen. De betreffende applicatie wordt dan zijn eigen data provider en dus dienen er additionele maatregelen genomen te worden om binnen die applicatie de UWV standaarden van op gebruiker gebaseerde authenticatie en autorisatie in te richten.

- Afwijkingen kunnen leiden tot architectuur afwijkingen en in bepaalde gevallen tot potentiele datalek meldingen

Gevolgen voor de DIM afnemers

UWV medewerkers bij de afnemende divisies krijgen alleen nog maar toegang tot de gegevens die ze voor hun werkzaamheden, taken en verantwoordelijkheden nodig hebben.

Er moet dus altijd sprake zijn van gegevensminimalisatie.

Het begint met geanonimiseerd tenzij …

Dan eventueel gemaskeerd

Als niet gemaskeerd worden bepaalde filters zoals het VIP filter toegepast tenzij …

Gevolgen voor de DWH medewerkers

Geen van de medewerkers in de DWH afdeling heeft toegang tot gegevens in Acceptatie en Productie tenzij er een incident is welke onderzocht moet worden. Hiervoor is er een envelop procedure ingericht zodat een DWH medewerker tijdelijk toegang krijgt op basis van een gerichte opdracht vanuit het management omdat er een incident is. (is als maatregel in de DIM GEB beschreven)

Meldplicht datalekken

Meldplicht datalekken | Autoriteit Persoonsgegevens

De meldplicht datalekken houdt in dat organisaties (zowel bedrijven als overheden) direct een melding moeten doen bij de Autoriteit Persoonsgegevens (AP) zodra zij een ernstig datalek hebben. En soms moeten zij het datalek ook melden aan de betrokkenen (de mensen van wie de persoonsgegevens zijn gelekt).

Wat is een datalek?

Bij een datalek gaat het om toegang tot of vernietiging, wijziging of vrijkomen van persoonsgegevens bij een organisatie, zonder dat dit de bedoeling is van deze organisatie.

Heb je een vermoeden dat er een datalek heeft plaatsgevonden meld dit bij je manager en bij de BSO van je divisie.


Vragen?