UITVOERING WERKNEMERSVERZEKERINGEN

Conceptueel ontwerp DIM

Project DataFabriek

Versie 0.7~~0.6~~

Datum: februari 2021~~december 2020~~

VERSIEBEHEER

Versie	Datum	Status	Korte beschrijving aanleiding / wijziging
0.2a	10/08/2020	Concept	Documentstructuur compleet; algemene zaken ingevuld, relevante onderdelen overgenomen uit bronzone-ontwerp (P17)
0.3	26/08/2020	Concept	Tussenversie
0.4b	02/09/2020	Concept	Voor interne review
0.5	14/09/2020	Concept	Interne review verwerkt, daarnaast cosmetische verbeteringen en een aantal uitbreidingen ter verduidelijking.
0.5a	14/09/2020	Concept	Nabranders interne review verwerkt. Ingebracht in TDA DWH
0.6	08/12/2020	Concept	Verwijzingen naar "doelbinding" consistent gemaakt met GEB. Commentaar architecten verwerkt. Ter beoordeling van het architectuur-deel (zie leeswijzer) weer aan hen opgeleverd.
0.7	18/02/2021	Concept	Bevindingen TDA DWH m.b.t. Gegevensvenster verwerkt

Versiebeheer: Gert-Jan Kooren
Opgesteld door: Gert-Jan Kooren, Richard Hogenberg, Diena Mahran, Ton Godtschalk, met input van overige leden van het projectteam.

Afgestemd met:	Status	Datum	Naam / Contactpersoon
TDA DWH		18/02/2021	Gerald van Leeuwen
DIV			Paula Kiens
Implementatieteam DWH/GD			Arnold Dikstaal

Besproken door:	Status	Datum
Project DataFabriek	Commentaar verwerkt in v0.5 / v0.5a	14/09/2020
Gerald/Ilse/John	Commentaar verwerkt in v0.6	Oktober 2020
Bas Marsman	Commentaar verwerkt in v0.6	Oktober 2020

Vertrouwelijkheid:
De lezer/gebruiker van dit document wordt geacht de inhoud daarvan vertrouwelijk te behandelen, tenzij uit de toelichting of bronvermelding blijkt dat de informatie als openbaar kan worden beschouwd.

INHOUDSOPGAVE

1 INLEIDING

1.1 Doel en inhoud van dit document

Dit conceptueel ontwerp beschrijft de opzet van het Data Integratie Magazijn (DIM).
Het werkt de grote lijnen uit de PSA DataFabriek verder uit, en dient als basis voor de detailontwerpen van de diverse bouwblokken binnen het DIM, zoals die worden ontworpen en gebouwd door het project DataFabriek

Het document beperkt zich daarom tot de delen van het DIM die in scope zijn voor dat project, en daarbinnen weer vooral op de scope van de fasen 4 en 5.[1]

Het conceptueel ontwerp is een levend document, en zal tijdens het project verder worden aangevuld en aangepast.

Deze versie van het document heeft in de administratie van het project DataFabriek productcode P177.

De eisen voor deze beheer- en systeemdocumentatie zijn, ten tijde van schrijven van dit ontwerp, nog niet helder, en worden nog besproken in en met werkgroepen van het Implementatieteam.

Dit conceptueel ontwerp mag dus niet (zoals in een eerder versie ten onrechte vermeld), beschouwd worden als de eerste stap naar de globale systeemdocumentatie. Het ontwerp levert daarvoor uiteraard wel belangrijke input.

Ten tijde van het schrijven van dit ontwerp wordt (buiten project DataFabriek of afdeling DWH) gewerkt aan de Doelarchitectuur Gegevensdiensten (DA GD), en aan de Enterprise Architectuur Gegevens-Huishouding (EA GH).
Beiden kunnen resulteren in wijzigingen in de PSA, en daarmee, na goedkeuring van die gewijzigde PSA en de impact er van op project-scope en –deliverables, ook in wijzigingen in dit ontwerp.

[1] Zie paragraaf 2.4 (DIM vs. project-scope) voor meer informatie hierover.

1.2 Leeswijzer

Dit conceptueel ontwerp valt uiteen in drie grote blokken:

- De hoofdstukken 2 t/m 6 beschrijven het DIM als geheel, en de algemene aanpak van (en eisen aan) de opslagstructuren en processen binnen de diverse zones.
- De hoofdstukken 7 t/m 13 beschrijven die opslagstructuren en processen per zone in meer detail
- De verdere hoofdstukken bieden meer detailinformatie (maar wel meestal relevant voor het hele DIM)
- De bijlagen bieden meer achtergrond op specifieke aspecten

De hoofdstukken 2 t/m 6 sluiten nauw aan op de PSA, en vallen qua kwaliteitsbewaking onder de architecten van project Datafabriek en afdeling DWH.
De overige hoofstukken en de bijlagen worden beheerd door de solution architecten en ontwerpers van het DIM. Kwaliteitsbewaking primair door ontwikkelaars (leesbaarheid, uitvoerbaarheid) en de TDA DWH (compliance met architectuur en andere ontwerpuitgangspunten).

Het DIM maakt gebruik van (concepten uit) algemene data warehouse methodieken (dimensioneel modelleren, DataVault2, etc.).
In dergelijke gevallen beschrijft dit ontwerp slechts het gebruik van deze methodieken binnen het DIM; de methodieken an sich worden hierbij bekend verondersteld.[2]

Notities voor de document-beheerder:

- *De meeste plaatjes in dit document zijn gemaakt met Visio. De originelen staan in het vsdx-bestand CO DIM (plaatjes).*
- *De overzichtsplaatjes van het DIM komen uit de PSA.*
- *Alle kruisverwijzingen in het document maken gebruik van de Word-velden voor dergelijke verwijzingen, zodat ze bij toevoegen van paragrafen/hoofdstukken eenvoudig ververst kunnen worden.*

[2] Paragraaf 1.3.3 (Achtergrondinformatie over methodieken) bevat een lijst met standaardwerken die hiervoor als achtergrondinformatie kunnen dienen.

1.3 Referentiedocumenten

1.3.1 Algemene referenties (UWV-standaarden en –beleid)

Hier alleen de belangrijkste. Zie de PSA voor een volledige lijst met relevante standaarden en beleidsdocumenten.

- Notitie Bewaarregels en Bewaartermijnen selectielijst UWV
- UWV ICT Richtlijn Data Anonimisering en Pseudonimisering Versie 1.1 24 juli 2018

1.3.2 Specifiek voor project DataFabriek [GL(1)]en/of DIM

- PSA DataFabriek, v2.0
- UWV HLD DIM, v1.0 (final), 2 september 2020
  N.B. Deze HLD is nog zonder de Fenced VLAN's
- DIM – Grondslagen verwerking persoonsgegevens, v1.0
- Typen gegevensgebruik door DIM, v1.0
- Interface-standaarden (bron-DIM), v1.4
  (werk in uitvoering; laatste werkversie in de werkmap van P182, laatste deelbare versie ook in de architectuur-folder)
- Datafabriek - Bescherming van persoonsgegevens in het DIM (v.0.5)
- Bescherming van persoonsgegevens in het DIM - Matrix Maskeringsklassen
- Metadata binnen Datafabriek
- Invulhulp RLO
- Standaarden modelleren en FO voor het DIM
- WoW Modelleren en FO maken in IDA

1.3.3 Achtergrondinformatie over methodieken

Dimensioneel modelleren:

- The Data Warehouse Toolkit
  (Ralph Kimball), ISBN 0471153370.
- Sterren en Dimensies
  (Harm van der Lek), ISBN 9-7894-92182180

DataVault2:

- Building a Scalable Data Warehouse with Data Vault 2.0
  (Dan Linstedt), EAN 9780128025109

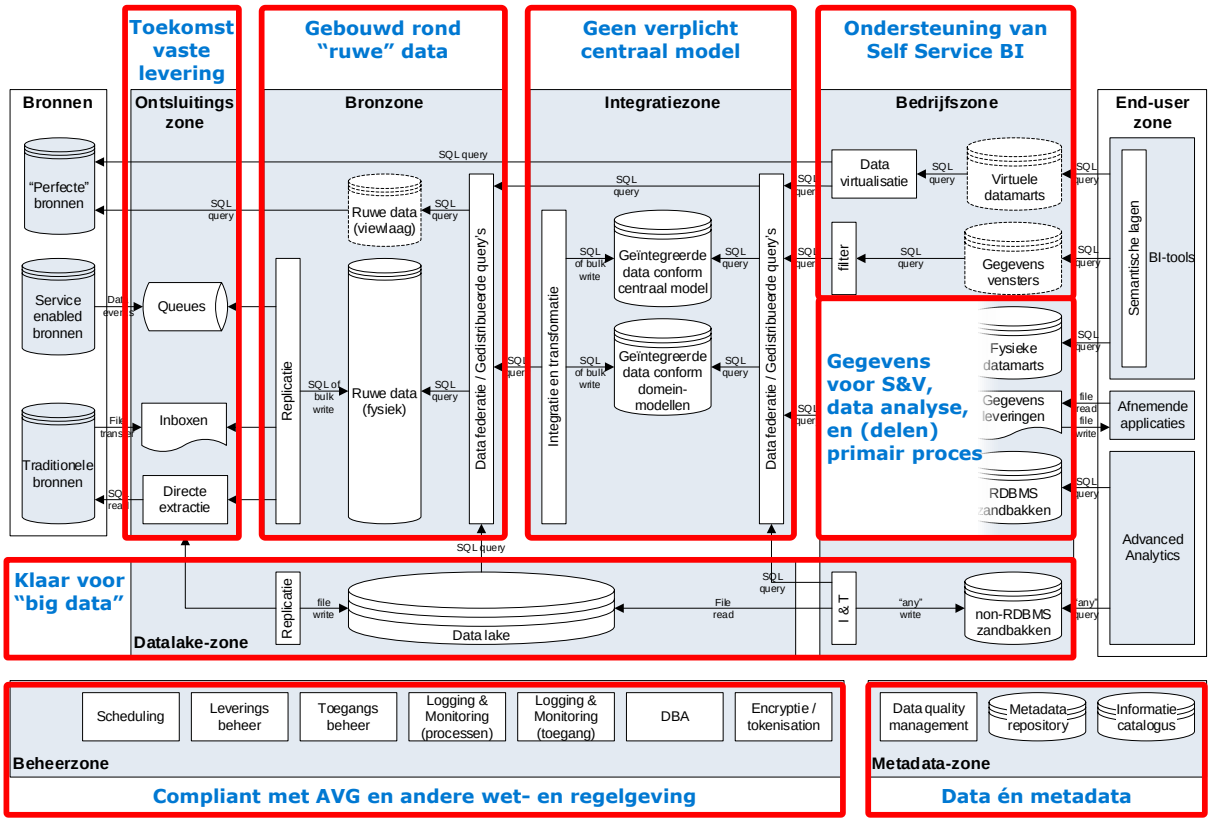
Logical Datawarehouse:

- Adopt the Logical Data Warehouse Architecture to Meet Your Modern Analytical Needs
  (Gartner / Henry Cook), Gartner ID: G00342254

2 HET DIM IN VOGELVLUCHT

2.1 Basisopzet DIM

Het Data Integratie Magazijn (DIM) vervangt de drie bestaande centrale data warehouses. Het is in essentie een "klassiek" enterprise datawarehouse (EDW), uitgebreid met een aantal "moderne" concepten die met name bedoeld zijn om de applicatie wendbaarder te maken (en zo het verkleinen van de time-to-market van nieuwe/gewijzigde informatievoorzieningen te ondersteunen), en om deze AVG-compliant te maken (door pseudonimisering/anonimisering en door gegevensminimalisatie). In een later stadium zal het DIM ook "big data" kunnen verwerken, maar dit valt niet binnen de scope van het project DataFabriek.



N.B. Het virtueel in de bronzone opnemen van "perfecte bronnen" is conceptueel mogelijk (vandaar dat ze zijn opgenomen in bovenstaand plaatje), maar dergelijke bronnen komen bij UWV nog niet voor.

2.2 Rol van het DIM binnen UWV

De rol van het Data Integratie Magazijn is vastgelegd in de PSA DataFabriek. Deze rol is:

- De verplichte leverancier (voor heel UWV) van gegevens t.b.v. verslaglegging[3] waarvoor UWV-brede consistentie vereist is;
- De verplichte leverancier (voor heel UWV) van divisie-overstijgende analysegegevens;
- De voorkeursleverancier van geïntegreerde/afgeleide gegevens (in bulk) met lage actualiteitseisen[4];
- De tweede voorkeursleverancier[5] van ruwe brongegevens (in bulk) met lage actualiteitseisen;
- Géén leverancier van "strikt vertrouwelijke" gegevens (met name medische en strafrechtelijke gegevens), tenzij er geen andere optie is;
- Géén leverancier van gegevens met hoge actualiteitseisen ("near realtime").
  Daar ligt het primaat bij service-oplossingen zoals die binnen het KIA/KOA-concept;
- Een mogelijke leverancier van alle overige gegevens.
  Er is geen beperking m.b.t. het type gegeven (anders dan de vertrouwelijkheid) of het type afnemend proces (anders dan indirect, via de actualiteitseisen)

Het Data Integratie Magazijn levert:

- Toegang tot de in het DIM opgeslagen[6] gegevens, bv. via een datamart of gegevensvenster, als input voor rapportage/analyse door tools buiten het DIM;
- Uitgaande interfaces / bestandsleveringen[7]: kopieën van in het DIM opgeslagen gegevens t.b.v. opslag/verwerking buiten het DIM;
- Metadata over al deze gegevens;
- Gegevensdefinities (technisch: data-elementen en functioneel: bedrijfstermen);
- Horizontale lineage (het pad dat een gegeven vanaf de bron tot aan de DIM-afnemer heeft doorlopen);
- Verticale lineage (de relatie tussen bedrijfsterm en de data-elementen waarin deze zijn terug te vinden);
- Datakwaliteitsinformatie[8].

Maar geen:

- Rapporten of dashboards.[9]

---

[3] En rapportages in het algemeen

[4] Het begrip "lage actualiteitseisen" is in de PSA (nog) niet uitgewerkt, maar is gerelateerd aan het voor het DIM haalbare service level op dit gebied. En dat is, voor de meeste brongegevens, een reguliere vertraging van één dag ("'s avonds geleverd is 's ochtends beschikbaar"), met daarbij de kanttekening dat dit bij incidenten (bij bron of DIM) kan oplopen tot enige dagen.

[5] De eerste voorkeur blijft de bronapplicatie zelf (als die daartoe in staat is)

[6] Of via het DIM virtueel ontsloten gegevens.

[7] In de PSA heet dit "gegevensleveringen". Omdat die term binnen DWH en Gegevensdiensten ook voor andere concepten wordt gebruikt, wordt in dit Conceptueel Ontwerp steeds de term "bestandsleveringen" gebruikt.

[8] Meestal gemaakt m.b.v. speciaal daarvoor verworven gespecialiseerde tools (Information Analyzer, QualityStage).
N.B. De verwerving en installatie (intussen afgerond) van deze tools was in scope van het project Datafabriek, maar de inzet ervan (nog) niet.

[9] De lijnorganisatie Datafabriek biedt wel een rapportagedienst o.b.v. van gegevens in het DIM, maar de daarvoor gebruikte BI/rapportage-tools vallen in de End-user zone, en daarmee buiten het DIM.

Ten tijde van het schrijven van dit ontwerp wordt (buiten project of afdeling DWH) gewerkt aan de Doelarchitectuur Gegevensdiensten (DA GD), en aan de Enterprise Architectuur Gegevens-Huishouding (EA GH).
Beiden kunnen resulteren in wijzigingen op de PSA, en dus ook, na goedkeuring van die gewijzigde PSA en de impact er van op project-scope en –deliverables, in wijzigingen op bovenstaande.

N.B. In een aantal gevallen valt de migratie van rapporten of dashboards wél in scope van het project Datafabriek.

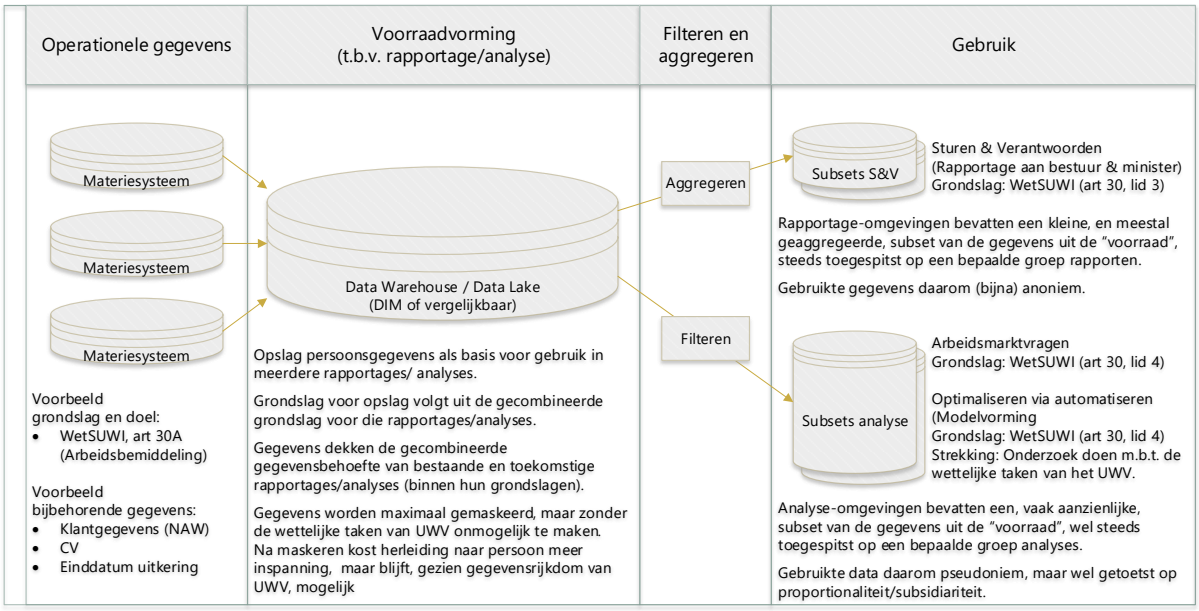
2.3 Gerelateerde Wet- en regelgeving

In Hoofdstuk 5 (Maskering~~Maskering~~) wordt de impact van wet- en regelgeving op het DIM in detail uitgewerkt.

2.3.1 Grondslagen verwerking persoonsgegevens

Juridische Zaken heeft, op verzoek van het project DataFabriek, onderzocht op welke wijze de voornaamste voor UWV geldende wet- en regelgeving (WetSUWI, Archiefwet, AVG) het DIM raken. De weerslag daarvan is beschreven in het document **DIM – Grondslagen verwerking persoonsgegevens, v1.0**.

Voor het DIM-ontwerp is onderstaand plaatje (overgenomen uit dat document) m.n. relevant:



2.3.2 Relevante bewaartermijnen

Voor het DIM zijn twee typen bewaartermijnen relevant:

- voor operationeel gegevensgebruik, typisch 5-7 jaar
  (zie **Notitie Bewaarregels en Bewaartermijnen selectielijst UWV**, bewaartermijnen voor Klantprocessen)

- voor analyse en rapportage, typisch 20-25 jaar
  (zie **Notitie Bewaarregels en Bewaartermijnen selectielijst UWV**, bewaartermijnen Ondersteunende processen, punt 15 en 16)

De eerste bewaartermijn valt onder de verantwoordelijkheid van de bron, de tweede onder die van (de eigenaar van) het DIM.

Voor operationeel gegevensgebruik kunnen gegevens niet (of nauwelijks) gemaskeerd worden, voor analyse/rapportage kan dat wel.

2.4 DIM vs. project-scope

xxx[GL(2]

3 ONTWERPUITGANGSPUNTEN

De PSA beschrijft 5 leidende principes voor de inrichting van de applicatiearchitectuur van het Data Integratie Magazijn:

1. Bewezen concepten
2. Geen realtime ambitie
3. Maximale ontkoppeling
4. Specifiek waar mogelijk, gemeenschappelijk waar noodzakelijk of waardevol
5. Kortste keten

Daarnaast volgen uit de drijfveren van het project nog 4 andere functionele eisen:

6. Compliant
7. Eenvoudig
8. Gebruiksvriendelijk
9. Gedefinieerd

Tenslotte zijn er nog 3 belangrijke, meer technische, eisen aan de gebouwde programmatuur:

10. Zoveel mogelijk RDBMS-onafhankelijk
11. Lineage mag niet gebroken worden
12. Minimale, en beheerbare, toolset

Deze 12 principes en eisen bepalen de randvoorwaarden waaraan ontwerp en implementatie van het DIM moeten voldoen. In andere woorden: de **ontwerpuitgangspunten**.

De precieze impact op en uitwerking van deze ontwerpuitgangspunten verschilt per DIM-zone. Zie daarvoor de diverse detail-hoofdstukken.

Ten tijde van het schrijven van dit ontwerp wordt (buiten project of afdeling DWH) gewerkt aan de Doelarchitectuur Gegevensdiensten (DA GD), en aan de Enterprise Architectuur Gegevens-Huishouding (EA GH). Beiden kunnen resulteren in wijzigingen op bovenstaande.

3.1 Bewezen concepten

De uitrol van het DIM moet zo voorspelbaar en beheersbaar mogelijk zijn. Opkomende gegevensintegratietechnologieën, zoals bv. data virtualisatie en Data Warehouse Automation (DWA) zijn daarom geen onderdeel van de "ruggengraat" van het DIM.[10]
Om ook stabiliteit en continuïteit op de langere termijn te kunnen garanderen dient het DIM overigens wel in de ruimte te voorzien om deze technologieën op een later tijdstip te incorporeren.

3.2 Geen realtime ambitie

Ook op de langere termijn wordt niet verwacht dat het DIM een rol gaat spelen in het leveren van actuele ("near realtime") gegevens.
Het DIM wordt daarom niet ingericht om (al dan niet op de langere termijn) gegevens met een hoge actualiteit te kunnen leveren, aangezien dit de oplossing nodeloos zou compliceren. Wel wordt, binnen het DIM, het aantal replica's geminimaliseerd (kortste keten principe) om zo de vertraging binnen het DIM te minimaliseren.

[10] Er worden overigens DWA-concepten toegepast, bv. in de ETL-harnassen.

Drie aandachtspunten daarbij:

a) Mocht een bron wél, middels een publish/subscribe-mechanisme (near) realtime willen leveren, dan zal het DIM dat ondersteunen. Binnen het DIM zullen de binnengekomen events dan worden verzameld om alsnog "in bulk" verwerkt te worden.
b) Bevat de bron historisch detail dat preciezer is dan de datum, dan zal het DIM dat detail gebruiken, door bijvoorbeeld, in de historie-opbouw (de versionering) meerdere wijzigingen per dag aan te kunnen. Dit geldt zowel wij meerdere leveringen per dag (zie punt hierboven) als bij minder frequente leveringen, met daarin meerdere wijzigingen per dag.
c) Er is geen beperking qua afnemend proces, mits dat proces kan leven met de beperkingen van het DIM (geen gegarandeerde performance/beschikbaarheid, beperkte actualiteit).

Het principe "geen realtime ambitie" mag dus niet gelezen worden als "geen rol in de dagelijkse operatie" of "alle historisch detail ten hoogste op dagbasis".

3.3 Maximale ontkoppeling

Om de flexibiliteit van de oplossing te maximaliseren worden de verschillende schakels in de door het DIM ondersteunde gegevensketen(s)[11] zoveel mogelijk ontkoppeld. Dit minimaliseert:

- de impact van wijzigingen in een bronsysteem op de opslag en verwerking van de gegevens uit dat bronsysteem in het DIM;
- de impact van wijzigingen binnen een component van het DIM op de overige componenten van dat DIM;
- de impact van wijzigingen binnen het DIM op de afnemers van dat DIM.

Ook binnen ETL-jobs worden de verschillende processtappen zoveel mogelijk ontkoppeld, bijvoorbeeld door modulair te ontwikkelen en optimaal gebruik te maken van, als "zwarte doos" aanroepbare, (herbruikbare) ETL-bouwstenen. Deze ontkoppeling dient twee doelen:

- verbeteren begrijpbaarheid van de ETL-jobs door opknippen in, elk op zich goed te begrijpen, deelfunctionaliteiten
- verbeteren wijzigbaarheid; een wijziging in een ETL-bouwsteen raakt de aanroepende ETL-job niet.

3.4 Specifiek waar mogelijk, gemeenschappelijk waar noodzakelijk of waardevol

Zowel in de positie van het DIM binnen de UWV enterprise architectuur (de "scope") als in de opzet van gegevensketens binnen het DIM zelf wordt, om de flexibiliteit van de oplossing te maximaliseren, géén "one size fits all" benadering toegepast.
In het ontwerp van het DIM vertaalt zich dat bijvoorbeeld in de "vraaggedreven" opzet van de integratiezone.[12]

3.5 Kortste keten

De time-to-market van de via het DIM geleverde gegevensproducten wordt verkort door toepassing van het "kortste keten" principe: voor een gegevensproduct worden nooit meer transformatie/replicatie-stappen uitgevoerd (of DIM-zones doorlopen) dan noodzakelijk om de functionele en non-functionele vereisten van een gegevensproduct te dekken.

Eigenlijk is "kortste keten" een verbijzondering van het ontwerpuitgangspunt "Specifiek waar mogelijk, gemeenschappelijk waar noodzakelijk of waardevol".

[11] Een "gegevensketen" is het geheel aan proces- en applicatie-stappen van gegevensbronnen naar (de tools en applicaties van) de afnemers

[12] Dat is iets anders dan "productgedreven"; zie paragraaf 9.2 (Informatiegebieden) voor meer details over de vraaggedreven opzet van de integratiezone

3.6 Compliant

Het DIM voldoet aan alle wet- en regelgeving (zie 2.3 Gerelateerde Wet- en regelgeving).

Het DIM voldoet grotendeels ook aan de ICT-principes, -richtlijnen en –standaarden. Waar dat niet zo is wordt die afwijking vooraf, en beargumenteerd, vastgelegd, en door de verantwoordelijke partijen goedgekeurd.

Voor het DIM zelf betekent compliance met wet- en regelgeving overigens vooral dat het de mogelijkheden moet bieden om de processen eromheen te helpen compliant te zijn.[13]
Denk aan ontwikkeling, testen, beheer, Data Lifecycle Management (DLM), doelbindings- en rechtsgrond-afspraken met de bronnen en de afnemers.

Compliance raakt allerlei aspecten van het DIM, w.o. metadata, maskering, bewaartermijnen, database-inrichting en toegangsbeheer.

Het ontwerpuitgangspunt "Compliant" kent, specifiek voor de AVG, twee verbijzonderingen:

- Privacy by design
- Privacy by default

3.6.1 Privacy by design

Artikel 25, lid 1 van de AVG stelt:
Rekening houdend met de stand van de techniek, de uitvoeringskosten, en de aard, de omvang, de context en het doel van de verwerking alsook met de qua waarschijnlijkheid en ernst uiteenlopende risico's voor de rechten en vrijheden van natuurlijke personen welke aan de verwerking zijn verbonden, treft de verwerkingsverantwoordelijke, zowel bij de bepaling van de verwerkingsmiddelen als bij de verwerking zelf, passende technische en organisatorische maatregelen, zoals pseudonimisering, die zijn opgesteld met als doel de gegevensbeschermingsbeginselen, zoals minimale gegevensverwerking, op een doeltreffende manier uit te voeren en de nodige waarborgen in de verwerking in te bouwen ter naleving van de voorschriften van deze verordening en ter bescherming van de rechten van de betrokkenen.

In bondig Engels: Privacy by design, oftewel: neem in je ontwerpen de bescherming van persoonsgegevens expliciet mee.
Dit wordt in het DIM onder andere ingevuld door:

- Al bij opslag maskeren van data
- Maximaal scheiden van gemaskeerde en ongemaskeerde data, zowel binnen DIM als bij doorlevering naar de afnemers
- Gescheiden DLM voor gegevens t.b.v. operationele processen en gegevens gebruikt t.b.v. analyse/rapportage
- Inzet van, bijvoorbeeld, gegevensvensters[14] om toegang-op-maat te ondersteunen
- Expliciete aandacht voor verwerking zeer gevoelige gegevens (Vertrouwelijkheidklasse 3)

[13] Twee voorbeelden daarvan:

1. Het DIM volgt, voor ongemaskeerde data, het Data Lifecycle Management (DLM) van de bron. Dat maakt het DIM an sich nog niet compliant; dat is pas het geval als de bron zelf zijn DLM compliant heeft ingericht. Wél zorgt de inrichting van het DIM ervoor dat áls de bron op dit gebied compliant is, het DIM dat ook is.
2. Het DIM kan gemaskeerde gegevens leveren aan afnemers. Dat maakt het voor die afnemers eenvoudiger om hun gegevensgebruik te laten voldoen aan de AVG-principes m.b.t. proportionaliteit en subsidiariteit.

) voor een uitwerking van de vraaggedreven aanpak.

[14] Zie paragraaf 14.5.3 (Gegevensvensters)

3.6.2 Privacy by default

Artikel 25, lid 2 van de AVG stelt:
De verwerkingsverantwoordelijke treft passende technische en organisatorische maatregelen om ervoor te zorgen dat in beginsel alleen persoonsgegevens worden verwerkt die noodzakelijk zijn voor elk specifiek doel van de verwerking. Die verplichting geldt voor de hoeveelheid verzamelde persoonsgegevens, de mate waarin zij worden verwerkt, de termijn waarvoor zij worden opgeslagen en de toegankelijkheid daarvan. Deze maatregelen zorgen met name ervoor dat persoonsgegevens in beginsel niet zonder menselijke tussenkomst voor een onbeperkt aantal natuurlijke personen toegankelijk worden gemaakt.

Samengevat: gebruik nooit meer of gevoeliger gegevens dan je nodig hebt.

Dit wordt in het DIM onder andere ingevuld door:
- Gemaskeerd leveren "tenzij", alleen leveren als afnemer rechtsgrond heeft én het gebruik voldoet aan de AVG-principes van proportionaliteit en subsidiariteit.
- Alleen opslag conform doelbinding/rechtgrond

3.7 Eenvoudig

Ontwikkel zo "eenvoudig" mogelijke oplossingen, primair gezien vanuit beheer–perspectief. Voorbeelden hiervan zijn:
- Maak oplossingen niet ingewikkelder dan noodzakelijk; bouw geen functionaliteit die (nog) niet nodig is (tenzij het later toevoegen veel aanpassingen en opnieuw testen vereist)
- Beheers de complexiteit en de begrijpelijkheid van de oplossingen door modulair ontwikkelen (w.o. ook inzet bouwstenen)
- Maak sowieso waar mogelijk gebruik van bouwstenen (bijv. log-bouwstenen). Houd daarbij wel in de gaten dat het niet zo mag zijn dat het herbruikbaar kunnen zijn van een bouwsteen niet weer leidt tot extra complexiteit binnen die bouwsteen.

Binnen dit ontwerpuitgangspunt moet je, met name voor generieke oplossingen, vaak een afweging maken. Het metadata-gedreven verwerken van bronleveringen, bijvoorbeeld, maakt de programmatuur zelf niet eenvoudiger, maar het gebruik er van weer wel.

Zie hoofdstuk 15 (Generieke oplossingen) voor een aantal handvatten voor deze afweging.

3.8 Gebruiksvriendelijk

Het DIM is uitdrukkelijk ingericht om self service gebruik te ondersteunen. Dat self service gebruik kan gebruik maken van gegevens uit alle zones van het DIM.[15]

Tegelijkertijd geldt echter het "kortste keten" principe; voor een gegevensproduct worden nooit meer transformatie/replicatie-stappen uitgevoerd (of DIM-zones doorlopen) dan noodzakelijk.

Vandaar dat, om onnodige nabewerking te voorkomen, voor alle z-ones (dus niet alleen de bedrijfszone) gebruiksvriendelijkheid voor de afnemer een aandachtspunt is:
- Begrijpelijke datamodellen
- Van elk gegeven is eenvoudig de functionele definitie te vinden
- Van elk gegeven is eenvoudig de oorsprong terug te vinden

3.9 Gedefinieerd

Alle gegevens in het DIM zijn volledig gedocumenteerd.

Alle ETL-processen binnen het DIM zijn volledig gedocumenteerd. Dit geldt ook voor hoe ze met data omgaan en wat er aan fout- en log-meldingen gegenereerd wordt.

[15] Wel met altijd tenminste een gegevensvenster ertussen

Een bronlevering wordt óf in zijn geheel verwerkt in de bronzone (inclusief gegevens met lage kwaliteit), óf, bij technische issues en/of gegevensfouten die de consistentie van het DIM bedreigen, geheel afgekeurd. Indien er bronleveringen worden afgekeurd moet dit worden vastgelegd en gelogd (geen uitval zonder foutboodschappen). Daarnaast dient er dan sprake te zijn van een actieve push naar DIM-beheer om de oorzaak te achterhalen en het probleem opgelost te krijgen.
Indien gegevens, om datakwaliteitsredenen, niet in een informatieproduct worden meegenomen, dan dient zulks helder in het FO en de gegevensdefinities van dat informatieproduct te zijn vastgelegd. Het heeft overigens de voorkeur om gegevens met lage kwaliteit als zodanig te markeren, i.p.v. ze niet mee te nemen.[16]

Geen "ongedocumenteerde" functionaliteit van standaard-tools inzetten; alleen gebruik maken van formeel beschreven functies van die standaard-software.

3.10 Zoveel mogelijk RDBMS-onafhankelijk

Het DIM is onderdeel van een landschap van analyseomgevingen met daarin ook andere dan Oracle-databases. Nu is dat alleen SQL Server, maar in de toekomst zullen daar ook NOSQL-oplossingen bijkomen. Tenslotte is het niet onmogelijk dat op lange termijn zelfs de DIM-databases ooit naar een andere RDBMS-oplossing verhuisd moeten worden.

Om ervoor te zorgen dat er zo min mogelijk RDBMS-specifieke afhankelijkheden ontstaan geldt het principe "DataStage/Optim, tenzij". Er moeten erg goede redenen zijn om van DataStage/Optim af te wijken. Deze redenen moeten geaccepteerd (door de IT-architect en/of de TDA[17]) en gedocumenteerd zijn.

Voorbeelden:

- Schrijf zelf geen SQL-code, maar gebruik de "optimize" functie van DataStage (die SQL genereert) als je een transformatie door de database (als ELT) wilt laten uitvoeren.
  Dit zorgt er niet alleen voor dat de transformatie met alleen hergenereren (dus zonder hercoderen) op een ander RDBMS (of zelfs op een NOSQL-platform) kan worden uitgevoerd, maar ook dat eenvoudig van transformatiepatroon (ETL of ELT) kan worden gewisseld als dat om, bijvoorbeeld, performance-redenen noodzakelijk blijkt.
- Gebruik OPTIM voor alle maskering, zodat de maskering ook in non-Oracle omgevingen kan worden gebruikt.

Bovenstaande betekent niet dat het gebruik van RDBMS-functionaliteit (of Oracle-specifieke functionaliteit) "verboden" is. Vaak wegen de voordelen ruim op tegen de nadelen. Ook het beschikbaar zijn van een eenvoudige manier om de functionaliteit te migreren naar een equivalent (een "exit-strategie") speelt hierbij mee.

3.11 Lineage mag niet gebroken worden

Om te zorgen dat de herkomst van ieder gegeven in een informatieproduct terug te voeren is naar de bron(nen) ervan moet er een ononderbroken horizontale lineage zijn van dat gegeven naar de attributen in de bron waaruit deze is opgebouwd.

In het algemeen wordt deze horizontale lineage automatisch aangemaakt door DataStage.

[16] De verwerving en installatie (intussen afgerond) van de hiervoor benodigde tools tools (Information Analyzer, QualityStage) was in scope van het project Datafabriek, maar de inzet ervan (nog) niet.
[17] Het precieze proces na in beheername moet nog worden bepaald

In twee gevallen maakt DataStage de horizontale lineage niet automatisch aan:

- **Generieke (metadata-gedreven) ETL-jobs**
  Bij generieke ETL-jobs worden de transformaties voor meerdere bronnen/doelen uitgevoerd door één enkele, metadata-gedreven, ETL-job; de zogenoemde "harnassen". DataStage maakt voor deze harnassen wel een horizontale lineage aan, maar deze is, juist door de generieke aard van de harnassen, weinig concreet.
- **Niet op DataStage gebaseerde transformaties/ontkoppelingen**
  In de context van het DIM betreft het hier vooral ontkoppelviews en gegevensvensters, en dan alleen als ze dusdanig complex zijn dat DataStage de lineage erdoorheen niet o.b.v. de database-metadata kan achterhalen.

Voor deze twee uitzonderingen moet een "workaround" worden toegepast om toch aan het ontwerpuitgangspunt te blijven voldoen. Deze workaround kan technisch zijn (via een omweg alsnog lineage genereren) of documentair (de lineage vastleggen in andere metadata en/of systeemdocumentatie).

Twee voorbeelden:

1. Het laden van data van de bronlevering tot en met de bronzone wordt uitgevoerd door metadata-gedreven harnassen. De resultaten daarvan worden ontsloten via de bronzone ontkoppelviews.
   Dit leidt ertoe dat er geen directe, eenduidig door DataStage gegenereerde, horizontale lineage is met de bronlevering.[18]
   Echter, doordat zowel de harnassen als de ontkoppelviews enkel technische transformaties uitvoeren (en dus geen businesslogica bevatten) kunnen de ontkoppelviews nog steeds beschouwd worden als een directe (één-op-één) representatie van de administratieve werkelijkheid van de bronnen. Voor de traceerbaarheid terug naar de bron voldoet dus een door DataStage gegenereerde horizontale lineage die begint bij de ontkoppelviews, gecombineerd met het opnemen van de bron (als technisch veld met metadata) in bron-zone en ontkoppelviews.[19]

   Mocht opname van de gegevensstroom door de harnassen en de ontkoppelviews in de toekomst toch noodzakelijk blijken, dan is dat waarschijnlijk mogelijk m.b.v. een "pseudo mapping" of een metadata-bridge. Inrichting daarvan is echter op dit moment nog geen onderdeel van de activiteiten voor 2021.

2. De gegevensvensters in de bedrijfszone bevatten een subset van de gegevens elders in het DIM. Deze gegevensvensters worden geïmplementeerd middels Oracle-views. DataStage houdt hier dus niet automatisch de horizontale lineage voor bij.
   Om toch de bronnen voor de inhoud van een gegevensvenster te kunnen bepalen zijn er twee opties[20]:
   - Het gegevensvenster altijd beperken tot "pure filtering", en die filters vastleggen in de functionele en technische documentatie van het venster.
     De horizontale lineage kan dan nét voor het gegevensvenster eindigen, omdat het venster zelf geen businesslogica bevat.

[18] Er is wel een indirecte horizontale lineage maar die is complex, omdat alle bronnen door een beperkt aantal processen, de zogenaamde harnassen, worden verwerkt.

[19] De lineage vanaf de ontkoppelview via de integratie- en bedrijfszone is wel gewaarborgd doordat de daarvoor gebruikte ETL niet generiek (metadata-gedreven) is, maar gebaseerd is op maatwerk-ETL.

[20] *Welke van de twee opties gebruikt gaat worden is afhankelijk van de mogelijkheden van de IBM-stack op dit gebied. Onderzoek hiernaar is onderdeel van Fase 5*

- Een technische workaround verzinnen die de gegevensvenster-views alsnog opneemt in de horizontale lineage, bijvoorbeeld door die zelf (gescript of gegenereerd) m.b.v. de standaard DataStage-functionaliteit hiervoor toe te voegen aan de lineage metadata.

  Uitbreiding van de lineage tot en met de gegevensvensters (en, indien nodig, vanaf de bronzone) is onderdeel van fase 5[GL(3].

3.12 Minimale, en beheerbare, toolset

De vereiste kennis voor nieuwe ontwikkelaars/beheerders, en daarmee hun leercurve, dient zo kort mogelijk te zijn.
Het DIM moet daarom gebaseerd zijn op een zo klein mogelijke toolset, en deze toolset moet, binnen UWV, goed ondersteund worden ("doeltechnologie" of vergelijkbaar).

Daarnaast heeft gebruik van binnen die tools beschikbare standaard-functionaliteit de voorkeur boven zelfbouw (binnen of buiten de tools).

Het minimale voor het DIM gebruikte toolset bestaat uit:

- Infosphere Information Server tools (DataStage, QualityStage, Optim, Information Analyzer, IGC)
- Oracle RDBMS en de daarbij behorende tools
- Shell scripts (bij voorkeur aangestuurd vanuit DataStage)
- IBM Workload Scheduler

4 TYPEN GEGEVENSGEBRUIK DOOR DIM

Dit hoofdstuk is gebaseerd op de presentatie Typen gegevensgebruik door DIM (1.0).

Het DIM gebruikt gegevens uit andere systemen voor drie verschillende doelen:

- als **basis** voor de informatiegebieden (combinaties van, al dan niet voorbewerkte, gegevens) waarop de informatieproducten (datamarts, leveringen, etc) zijn gebaseerd
- als **parameters** bij het creëren van die informatieproducten
- als input voor de **beveiliging** van informatieproducten

Elk van deze gebruiksdoelen stelt eigen eisen aan de wijze waarop de gegevens aan het DIM ter beschikking worden gesteld.

4.1 Gegevens als basis – brede bronontsluiting

Het overgrote deel van de gegevens in het DIM wordt gebruikt als basis voor informatieproducten. Deze gegevens worden door de bron aangeleverd middels een gestandaardiseerde bronlevering. De regels hiervoor zijn beschreven in het document **Interface-standaarden (bron-DIM)**.
De meeste van deze bronontsluitingen zijn "brede bronontsluitingen", ze dekken alle brongegevens waarvoor een bestaande óf een verwachte behoefte bestaat bij één of meer van de DIM-afnemers (zieRationale brede bronontsluiting).

Smal	Breed	Volledig
De bronlevering bevat alleen de brongegevens waaraan een **expliciete** behoefte is (bij de gegevensafnemers).	De bronlevering bevat alleen de brongegevens waaraan een **verwachte** behoefte is (bij de gegevensafnemers).	De bronlevering bevat de volledige inhoud van het bronsysteem.

De term "breed" kan voor de meeste bronnen daarbij gelezen worden als:
"alle objecten met een bestaande behoefte, en daarbinnen alle attributen (ook die zonder bestaande behoefte), tenzij deze attributen geen relevantie hebben buiten het bronsysteem".

Voor een beperkt aantal bronnen wordt, al dan niet tijdelijk, slechts een smalle bronontsluiting geïmplementeerd.
Dit is alleen toegestaan als brede bronuitsluiting op compliance-problemen stuit (bijvoorbeeld omdat de gegevens eigendom zijn van een partij buiten UWV ligt, of omdat het HRM-data betreft), of als het direct implementeren van een brede bronontsluiting leidt tot onoverkomelijke planningsissues bij de bron.

N.B. Het doorleveren van persoonsgegevens uit de bron aan het DIM (om het daar ook op te slaan) is an sich geen AVG-issue; het enige dat er immers gebeurt is dat de dezelfde data nu op twee plekken staat. Voor de AVG hoeft de levering dus niet versmald te worden.
Daarnaast is levering van een gegeven aan het DIM géén vrijbrief voor het doorleveren van die gegevens aan de afnemers van dat DIM: die moeten daarvoor nog steeds aantonen dat het gebruik van de gegevens rechtsgrond heeft , en voldoet aan de AVG-principes van ~~van~~ proportionaliteit en subsidiariteit.

4.1.1 Belang en rol RLO

Voor gegevens die als basis worden gebruikt, is nauwkeurige metadata van groot belang:

- De gegevens in het DIM moeten gebruikt kunnen worden bij het maken van informatieproducten (door de Datafabriek en/of, bij self service, door de afnemers van het

DIM) zonder dat hiervoor navraag nodig is bij de bronnen, en, vooral, zonder dat hiervoor aannames t.a.v. de betekenis moeten worden gedaan.
- Eenduidige gegevensdefinities en helder gedefinieerde vertrouwelijkheid (en andere gebruiksbeperkingen) zijn een randvoorwaarde voor compliance; als de betekenis van een gegeven niet duidelijk is kan het vereiste beveiligings- en gebruiksregime (doelbinding, proportionaliteit, etc.) niet bepaald worden. Ook het maskeren van identificerende attributen vereist eenduidige gegevensdefinities.
- Historieopbouw van de brongegevens (in de DIM-bronzone) kan alleen betrouwbaar worden uitgevoerd als het historisch gedrag van die gegevens (in de bron) bekend is.

Al deze metadata wordt (door de bron-beheerders en/of eigenaars) vastgelegd in een RLO (Record LayOut). Deze RLO is de basis voor de inrichting van de verwerking van de brongegevens binnen het DIM. De RLO is daarmee ook integraal onderdeel van de leverafspraken tussen bron en DIM (zie verder **Interface-standaarden (bron-DIM)**).

N.B. Als de bron beschikt over een actueel FUGEM en TEGEM, dan vereenvoudigt dit het opstellen van een RLO aanzienlijk.

4.1.2 Geen weigering aan de poort

Bronleveringen worden alleen, in hun geheel, afgewezen (en dus niet verwerkt in het DIM) als de levering (of de inhoud ervan) technisch niet verwerkbaar is.

Gegevens met kwaliteitsissues (lege verplichte velden, loshangende verwijssleutels, etc) worden gewoon geladen. Die lage kwaliteit is immers de administratieve werkelijkheid van de bron.

N.B. Kwaliteitsissues die de verwerkbaarheid raken (lege sleutelvelden of dubbele primaire sleutels, bijvoorbeeld), resulteren het afwijzen van de volledige levering.

4.1.3 Zeer gevoelige gegevens

Binnen het DIM en de daaraan gerelateerde vraagsturings- en beheerprocessen bestaat expliciete aandacht voor privacy-risico's.
Voor gegevens met vertrouwelijkheidsklasse 3 (VK3) gelden een aantal extra regels:
- In principe laadt het DIM géén VK3-gegevens[21].
- Als er een zéér zwaarwegend belang is om zulks toch te doen (dit ter beoordeling van bron én afnemer, en vergezeld van een GEB[22]), dan kunnen VK3-gegevens toch binnen het DIM worden verwerkt. Verwerking zelf is hierbij "DIM-standaard", maar opslag van de gegevens, en autorisatie voor de toegang daartoe, is gescheiden van die voor de overige gegevens.
- Een bijzonderheid zijn gegevens die alleen VK3 hebben voor bijzondere personen/zaken, of bijzondere waarden. Deze worden in het algemeen wél ingelezen in het DIM, maar alleen voor zeer specifiek toegestaan gebruik ter beschikking gesteld aan afnemers. Zie volgende twee paragrafen voor meer details.

4.1.3.1 Zeer gevoelige personen/zaken

In een aantal gevallen worden gegevens pas gevoelig als ze betrekking hebben op een specifieke persoon of zaak:

- **VIP's**
  Als iemand de VIP-status krijgt (of zijn VIP-status verliest), dan geldt dat met terugwerkende kracht.

[21] Zie paragraaf 2.2 (Rol van het DIM binnen UWV)

[22] GegevensbeschermingsEffectBeoordeling (Engels: Privacy Impact Assessment; PIA)

Daarnaast moeten gegevens van VIP's wel worden opgenomen in diverse totaaltellingen t.b.v. Sturen en Verantwoorden (S&V).
Vandaar dat VIP-gegevens wél aan het DIM moeten worden geleverd, ook al zijn ze VK3.
Binnen het DIM worden deze gegevens vervolgens afgeschermd van de overige gegevens door toegang ertoe altijd te valideren tegen de laatste versie van de "VIP-lijst" uit UPA.[23]

- **Bijzondere Zaken (BZ), Bijzondere GevalsBehandeling (BGB), EP+ (Eigen Personeel + familie)**
  UWV kent op dit moment geen centrale registratie van BZ, BGB en EP+; classificatie als zodanig verschilt per bron.[24] Daarnaast moeten gegevens betreffende dergelijke zaken/personen wel worden opgenomen in diverse totaaltellingen t.b.v. S&V.
  Vandaar dat deze gegevens wél aan het DIM moeten worden geleverd, ook al zijn ze, deels, VK3.
  Binnen het DIM worden deze gegevens vervolgens, waar nodig en mogelijk, afgeschermd van de overige gegevens middels bron-specifieke filters op de informatieproducten. De functionaliteit van dergelijke filters moet door de bron worden gevalideerd.

4.1.3.2 Velden met zeer gevoelige waarden

Soms zijn alleen bepaalde waarden in een veld "zeer gevoelig", en is dat voor de meerderheid van de waarden niet het geval:

- **Velden met vrije tekst**
  In vrije tekstvelden kunnen allerlei gegevens staan, dus ook, in theorie, zeer gevoelige (VK3) gegevens.
  In de RLO dient te worden aangeven hoe "vrij" een dergelijk tekstveld is, en of de kans op zeer gevoelige gegevens in dat veld reëel is. Is zulks het geval, dan wordt het veld behandeld als VK3.

  Velden met vrije tekst moeten in het algemeen wél aan het DIM geleverd worden. Dit, bijvoorbeeld, voor datakwaliteitsanalyse. Uitzonderingen zijn velden die geen enkele waarde hebben buiten de bron zelf, en velden waarvan het belang voor rapportage/analyse te klein is om het risico op VK3-gegevens te rechtvaardigen.

- **Gevoelige waarden in gestructureerde velden (velden met een vastgesteld waardebereik)**
  Als bepaalde waarden in een gestructureerd veld VK3 zijn (bv. beroep=sekswerker, reden stopzetting uitkering = detentie), dan worden deze waarden in de informatieproducten (ook de ongemaskeerde) onherkenbaar gemaakt, bijvoorbeeld door deze te vervangen door de waarde "overig" (of equivalent).
  In het DIM zelf zijn ze wel herkenbaar opgeslagen, omdat het mogelijk moet blijven om, bij voorbeeld, te kunnen berekenen hoe groot het aantal stopgezette uitkeringen is dat "detentie" als reden heeft.
  Het is ook hierom dat gestructureerde velden met mogelijke VK3-waarden vrijwel altijd wél aan het DIM geleverd moeten worden.

4.2 Gegevens als parameters

Parameters worden gebruikt om het aanmaken van informatieproducten te sturen, en niet als basis voor analyse/rapportage. Ze worden pas opgehaald als ze nodig zijn (tijdens het aanmaken van

[23] Zie paragraaf 4.3 (Gegevens t.b.v. beveiligingGegevens t.b.v. beveiliging)

[24] Er bestaan plannen om binnen UWV een centrale registratie voor BZ/BGB/EP+ op te zetten.
Als er in de toekomst een centrale registratie voor BZ/BGB/EP beschikbaar komt, dan zal het DIM erop gaan aansluiten.

het betreffende informatieproduct) en worden alleen in het DIM opgeslagen als zulks voor de traceerbaarheid of reproduceerbaarheid van het informatieproduct vereist is.

Het DIM onderscheidt twee typen parameters:

- **Abonnementsgegevens**
  Soms moeten informatieproducten (m.n. bestandsleveringen) samengesteld worden o.b.v. een "abonnement", bijvoorbeeld zoals opgeslagen in UGC.
  Deze abonnementsgegevens worden pas bij het aanmaken van het informatieproduct door het DIM opgehaald (bv. via een service). Ze worden dus niet, zoals bij de basis-gegevens, middels een bronlevering vergaard. De via de service opgehaalde parameters worden wel in het DIM vastgelegd, maar dat is t.b.v. de traceerbaarheid/reproduceerbaarheid van het erop gebaseerde informatieproduct, en **niet** t.b.v. verdere analyse. [25]
  N.B. Levering van abonnementsgegevens als parameters is volledig gescheiden van levering van abonnementsgegevens als basis-gegevens (t.b.v. analyse/rapportage over de abonnementen zelf).
  Als er, bijvoorbeeld, in de toekomst een behoefte ontstaat om via het DIM het abonnementsgedrag van abonnees te analyseren, dan zal UGC dan op twee manieren gegevens aan het DIM leveren: als "gewone" bron (via een brede bronontsluiting), en als provider van de services met parameters.

- **Ad hoc parameters**
  Gebruik van ad hoc parameters zal in het algemeen vooral voorkomen in de End-userzone, en dus bij het uitlezen van een informatiegebied, en niet bij het aanmaken ervan. Deze parameters vallen daarmee buiten het DIM.

4.3 Gegevens t.b.v. beveiliging

Gegevens betreffende gevoelige personen/zaken worden wél in het DIM opgeslagen, maar toegang ertoe is afgeschermd.[26]

Voor deze afscherming zijn "beveiligingsgegevens" nodig.

Op dit moment haalt het DIM maar één type beveiligingsgegevens op:

- **VIP's**
  Voor VIP's is deze afscherming gebaseerd op de "VIP-lijst" uit UPA[27]. Levering van deze gegevens aan het DIM volgt technisch dezelfde route als die voor basis-gegevens, maar heeft uitdrukkelijk alleen beveiliging als doel, en is dus géén brede bronontsluiting.
  Opslag van de VIP-lijst is dan ook gescheiden van opslag van "gewone" brongegevens.
  N.B. Mochten, in de toekomst, UPA-gegevens ook gebruikt gaan worden in rapportages/analyses, dan zal UPA op twee manieren gegevens aan het DIM gaan leveren: als "gewone" bron (via een brede bronontsluiting), en als leverancier van beveiligingsgegevens.
  Dit omdat de twee leveringen een volledig ander doel hebben, en dus ook anders beargumenteerd en bestuurd moeten worden.

Beveiligingsgegevens worden binnen het DIM op dezelfde wijze verwerkt als gewone (brede) bronontsluitingen (zie paragraaf 4.1 (Gegevens als basis – brede bronontsluiting), zij het dat er in

[25] Voor fase 4/5 worden alleen voor de SUAG-levering abonnementsgegevens gebruikt. Deze worden, via standaardservices, opgehaald uit UGC en POLIS+

[26] Zie paragraaf 4.1.3.1~~4.1.3~~ (Zeer gevoelige personen/zaken~~Zeer gevoelige personen/zaken~~)

[27] Een officiële lijst met alle BSN's die als VIP moeten worden beschouwd. Het betreft hier "echte VIP's", dus niet eigen personeel o.i.d.

de gebruiksbeperkingen (in RLO en/of GLO) expliciet is opgenomen dat de betreffende gegevens alleen voor beveiligingsdoeleinden mogen worden gebruikt (en dus niet voor analyse op, bijvoorbeeld, het VIP-zijn zelf).

Er bestaan plannen om binnen UWV een centrale registratie voor BZ/BGB/EP+ op te zetten. Zodra die er is zal het DIM erop gaan aansluiten. Deze aansluiting (en de opslag van de via die aansluiting vergaarde gegevens) zal vergelijkbaar zijn met die voor VIP's.
Tot die tijd zal het afschermen van BZ/BGB/EP+ gebaseerd zijn op, via brede bronlevering verkregen, gegevens uit de bron zelf.[28]

[28] Zie 4.1.3.14~~.1.3~~ (Zeer gevoelige personen/zaken~~Zeer gevoelige personen/zaken~~)

5 MASKERING EN DLM

Zoals beschreven in paragraaf 2.3 (Gerelateerde Wet- en regelgeving) heeft het DIM te maken met twee typen gebruik (rapportage/analyse en meer operationeel) en twee typen bewaartermijnen.

Het DIM ondersteunt daarom een "meersporenbeleid" m.b.t. gegevensmaskering/filtering en Data Lifecycle Management (DLM).

5.1 Minimalisatie privacy-risico's in 3 stappen

Stap 1: Filter bij de bron

- In principe laadt het DIM géén VK3-gegevens (zie PSA DataFabriek, "Rol van het DIM binnen UWV"), tenzij de VK van die gegevens over tijd kan wijzigen (bv. VIP's)

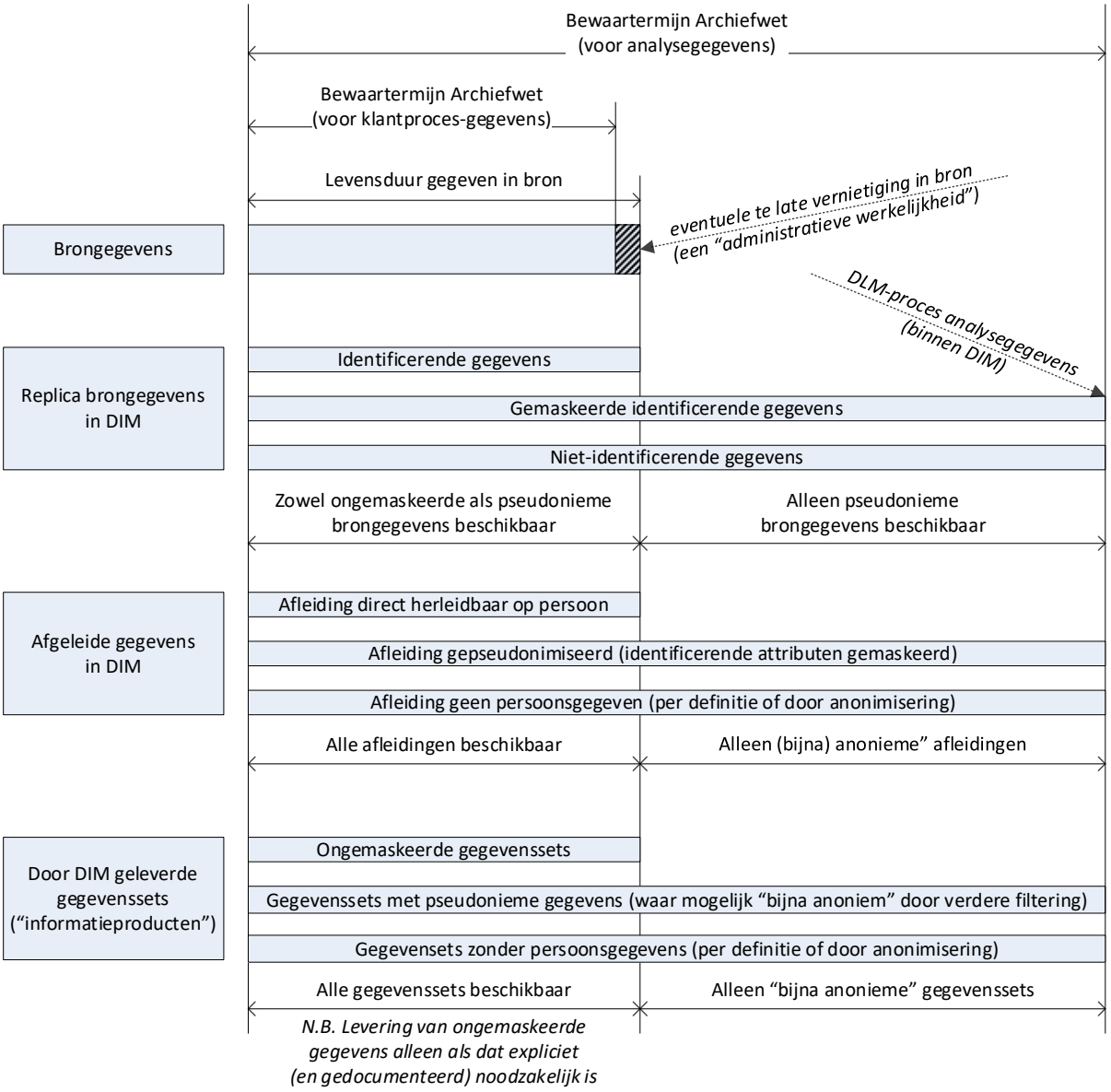
Stap 2: Maskering bij opslag

- In de bronzone van het DIM worden de gegevens uit de bron zowel gemaskeerd als ongemaskeerd opgeslagen. In de gemaskeerde versie zijn alle identificerende attributen zoveel mogelijk onherkenbaar gemaakt. Dit binnen de randvoorwaarde dat de gemaskeerde gegevens zinnig en gedetailleerd genoeg moeten blijven voor elke door het DIM ondersteunde rapportage en/of analyse.
- De ongemaskeerde gegevens worden na het verstrijken van de "operationele" bewaartermijn verwijderd, de gemaskeerde gegevens na het verstrijken van de bewaartermijn voor analyse- en rapportageprocessen.

Stap 3: Gegevensminimalisatie bij levering

- De informatieproducten uit het DIM zijn "by default" gebaseerd op de gemaskeerde gegevens.
- De informatieproducten bevatten filters op tabel-, rij- en attribuutniveau. Dit om te garanderen dat de getoonde deelverzameling alleen gegevens bevat waarvoor rechtsgrond geldt, en waarvan het gebruik voldoet aan de AVG-principes m.b.t. proportionaliteit en subsidiariteit
- Waar mogelijk worden gegevens ook geaggregeerd, en op die wijze geanonimiseerd.

5.2 Twee typen Data Lifecycle Management



Het DLM in het DIM volgt twee sporen:

- Gegevens die binnen UWV om operationele redenen beschikbaar zijn, worden in ongemaskeerde vorm ook in het DIM vastgelegd.
  De bewaartermijn van die gerepliceerde gegevens volgt daarbij de bewaartermijn van de originele gegevens. De bron blijft verantwoordelijk voor het uitvoeren van het Data Lifecycle Management (DLM) op die originele gegevens[29]; het DIM volgt slechts en is daarmee "in commissie niet compliant" als een bron een gegeven te vroeg of te laat verwijdert.
- Gegevens die binnen UWV, na verloop van de operationele bewaartermijn, beschikbaar dienen te blijven t.b.v. rapportage en analyse blijven conform die langere bewaartermijn[30] beschikbaar in het DIM, zij het alleen in gemaskeerde vorm. Gezien de gezamenlijke detailbehoefte van de achterliggende rapportage- en analyseprocessen blijft deze data (indirect) herleidbaar naar personen, en moet dus beschouwd worden als pseudoniem.

[29] Conform de hiervoor relevante bewaartermijnen uit de Selectielijst, typisch 5-7 jaar

[30] Typisch 20-25 jaar, zie Selectielijst onder "Bewaartermijnen voor ondersteunende processen, punt 15 en 16

Voor alle gegevens geldt dat ze uiteindelijk uit het DIM worden verwijderd. Dit om compliant te blijven met de Archiefwet.

De algemene aanpak en de functionele behoeften van zowel maskering als DLM staan beschreven in het overkoepelende document **Datafabriek - Bescherming van persoonsgegevens in het DIM** en het detailspreadsheet **Bescherming van persoonsgegevens in het DIM - Matrix Maskeringsklassen**.

5.3 Twee typen DLM impact op historische consistentie

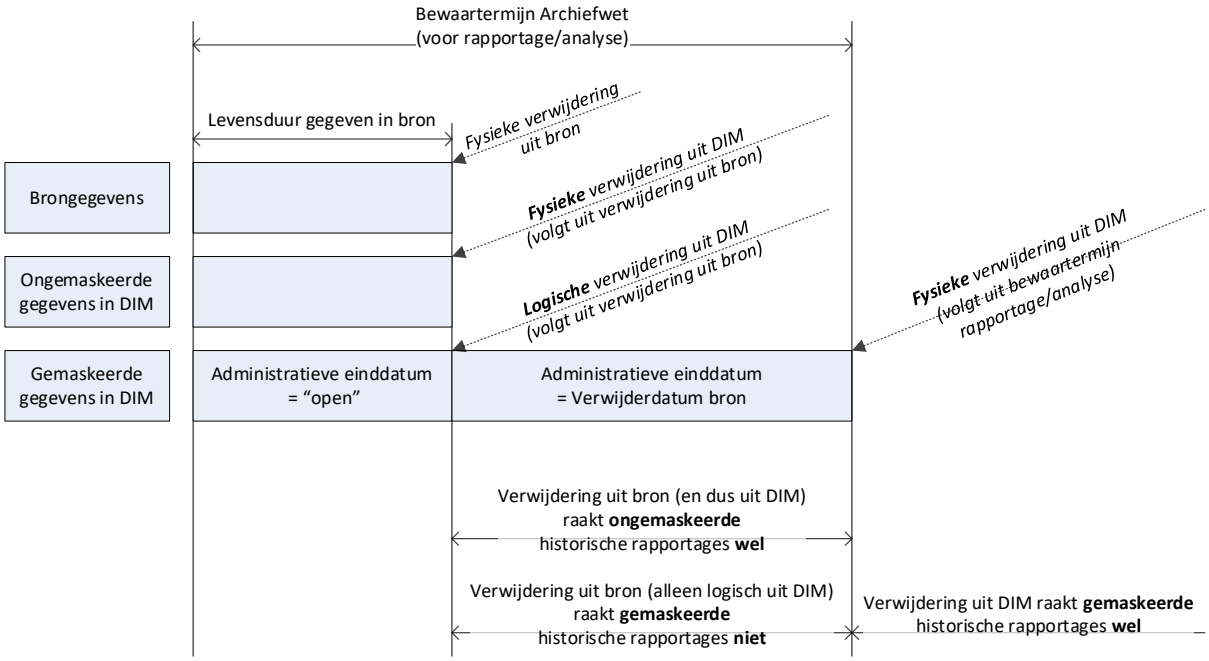
De bewaartermijn van gemaskeerde gegevens is langer dan die van ongemaskeerde gegevens. In de gemaskeerde gegevens is wél te herkennen wanneer het originele gegeven uit de bron (en dus ook uit de ongemaskeerde gegevens is verwijderd: dat gegeven krijgt dan in de gemaskeerde versie een "administratieve einddatum". Dit heet in de DWH-theorie "logisch verwijderen".

Het fysiek verwijderen van een gegeven uit de bron heeft, omdat alle administratieve versies dan uit de ongemaskeerde gegevens het gegeven dan ook uit fysiek het DIM verwijderd worden, heeft **impact** op **ongemaskeerde** historische rapportages waarin dat gegeven is opgenomen; bij opnieuw draaien zullen die rapportages andere getallen laten zien.

Het fysiek verwijderen van een gegeven uit de bron heeft **geen impact** op **gemaskeerde** historische rapportages waarin dat gegeven is opgenomen, aangezien het DIM daar slechts logisch verwijderd; bij opnieuw draaien zullen die rapportages de oorspronkelijk gerapporteerde getallen laten zien (mits het rapport gebruik maakt van een expliciete administratieve peildatum).

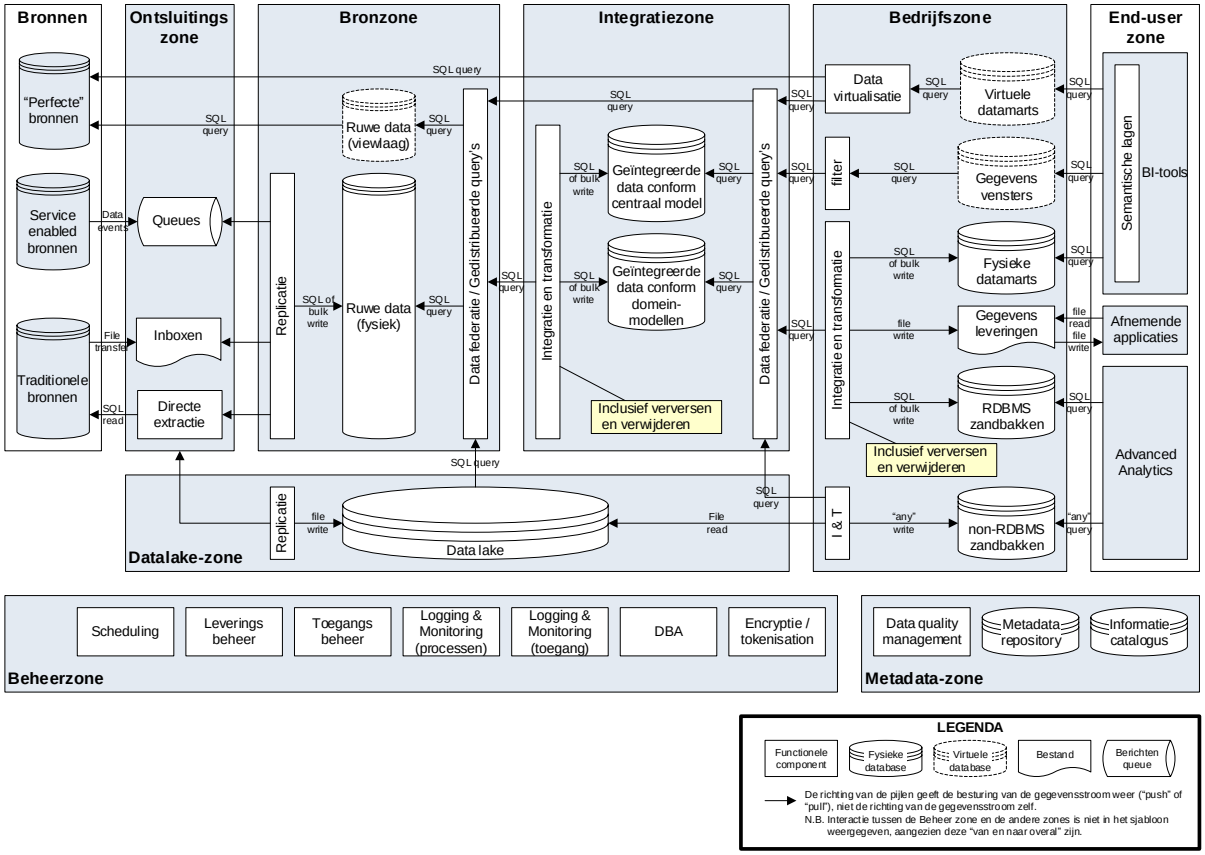
Het fysiek verwijderen van een gemaskeerd gegeven uit het DIM omdat ook de langere bewaartermijn voor analyse/rapportage is verstreken heeft **wel** impact op **gemaskeerde** historische rapportages waarin dat gegeven is opgenomen, ook als die rapportages gebruik maken van een expliciete administratieve peildatum.

Waar volledige reproduceerbaarheid van rapporten cruciaal is, dient het rapport dus gebaseerd te zijn op apart opgeslagen afgeleide gegevens (gemaskeerd en/of geaggregeerd), waarvan de input (zelf weer m.b.v. een administratieve peildatum is vergaard).



6 VOORNAAMSTE DIM-ONDERDELEN

6.1 De DIM-zones



De kern van het Data Integratie Magazijn bestaat uit 4 zones:

- De **Ontsluitingszone** zorgt voor de technische ontkoppeling tussen bronsystemen en DIM;
- De **Bronzone** "compenseert" de eventuele tekortkomingen van de bronsystemen m.b.t. gegevensbeschikbaarheid en historisch besef;
- De **Integratiezone** bevat, t.b.v. bruikbaarheid, consistentie en efficiëntie, veel (en/of verplicht) gebruikte gegevensmodellen en afgeleide gegevens;
- De **Bedrijfszone** stelt de data in het DIM "op maat" beschikbaar aan de diverse eindgebruikers (mensen en machines).

Bovenstaande vier zones vormen het hart van de gegevensketen van de gegevensbronnen naar (de tools en applicaties van) de afnemers in de, buiten het DIM vallende, **End-user zone**. Ze zijn grotendeels gebaseerd op RDBMS-technologie, en daarmee vooral geschikt voor verwerking van relationeel gestructureerde en qua structuur redelijk stabiele data.

Voor minder/anders gestructureerde, of qua structuur sneller wijzigende, gegevens bevat het DIM een **Datalake-zone**.

De Datalake-zone is niet in scope van het project DataFabriek, en wordt in dit document verder niet beschreven.

De technische en business metadata betreffende de (verwerking van) gegevens in het DIM zijn

terug te vinden in de **Metadata-zone**. De metadata repository en de informatiecatalogus in deze zone bevatten alle technische respectievelijk functionele metadata die noodzakelijk is voor gegevensverwerking en -levering door het DIM. Deels wordt deze metadata door het DIM zelf gegenereerd (m.n. horizontale lineage), deels wordt deze overgenomen uit elders gecreëerde metadata (m.n. functionele gegevensdefinities brondata uit de FUGEMs). De koppeling tussen technische velden en hun functionele equivalent (de verticale lineage) zal ook binnen de DIM-metadata worden vastgelegd.

De functionaliteiten t.b.v. besturing en beheer van het DIM, ten slotte, zijn gegroepeerd in de **Beheer zone**.

6.2 Koppelvlakken

6.2.1 Koppelvlakken tussen de RDBMS-zones

Tussen de 4 kernzones van het DIM, en tussen de leveranciers en afnemers van die zones, liggen koppelvlakken:

- Tussen de **Bronnen** en de **Ontsluitingszone** liggen de **bronleveringen** (in diverse technische formaten);
- Tussen de **Ontsluitingszone** en de **Bronzone** ligt de **staging laag**;
- Tussen de **Bronzone** en de achterliggende zones (**Integratiezone** en **Bedrijfszone**[31]) liggen de **bronzone-ontkoppelviews**;
- De **informatieproducten** in de **Bedrijfszone** zijn eigenlijk zelf het koppelvlak tussen de voorliggende zones en de **End-user zone.**

Voor het aanmaken van deze **informatieproducten** is, in sommige gevallen, een koppelvlak voor het ophalen van parameters (m.n. abonnementsgegevens) vereist.

6.2.2 Koppelvlakken met de Metadata-zone

Alle kernzones, en deels ook de koppelvlakken daartussen, interacteren in één of andere vorm met de **Metadata**-zone:

a) Ze worden gedreven door stuur-metadata;
b) Ze leveren log-metadata op;
c) Ze vallen onder datakwaliteitsmanagement[32];

Daarnaast krijgen eindgebruikers ook direct toegang tot metadata (bijvoorbeeld in de **Informatiecatalogus**).

De metadata-koppelvlakken onder (a) en (b) worden beschreven in de detailhoofdstukken per zone en/of koppelvlak. Datakwaliteitsmanagement wordt beschreven in Bijlage G: Datakwaliteitsbeheer, en directe toegang tot metadata (door eindgebruikers) wordt beschreven in hoofdstuk 12 (Metadata-zone).

6.2.3 Koppelvlakken met de Beheerzone

*De koppelvlakken met de **beheerzone** worden door project en lijn gezamenlijk uitgewerkt, als onderdeel van het in beheer nemen van de OTAP-straat door de afdeling DWH.*
De uitkomsten van die uitwerking zullen in een latere versie van dit ontwerp worden verwerkt.

[31] Vanwege het "kortste keten" principe kan een informatieproduct in de Bedrijfszone direct koppelen met de Bronzone (dus zonder een tussenstap door de Integratiezone)
[32] Hier voor de volledigheid opgenomen: Inrichting van datakwaliteitsmanagement valt (nog)niet in scope van het project Datafabriek.

6.2.4 Koppelvlakken met de Datalake-zone

Deze koppelvlakken zijn dit document nog niet beschreven.
De Datalake-zone valt buiten scope van het project, en de rol en technologie ervan zijn nog onduidelijk, en afhankelijk van beslissingen elders in UWV.

6.3 Karakteristieken van de gegevens in de vier kern-zones

De **ontsluitingszone** is primair technisch van aard. De gegevens in deze zone zijn niet relevant (en dus ook niet toegankelijk) voor eindgebruikers.

In zowel **ontsluitingszone** als **bronzone** is de verwerking in hoge mate gestandaardiseerd. Dit maakt het mogelijk om deze verwerking agile en robuust te maken door het toepassen van concepten uit Data Warehouse Automation (DWA). De ETL-harnassen zijn hiervan het duidelijkste voorbeeld.

De **bronzone** is het hart van het DIM. Hier wordt de administratieve werkelijkheid van de DIM-bronnen opgeslagen en, waar nodig, voorzien van extra historie, bijvoorbeeld door een langere bewaartermijn dan de bron aan te houden (langere historie) of door het afleiden van de administratieve tijdlijn uit de verschillen tussen opeenvolgende bronleveringen (gedetailleerdere historie).
Daarnaast wordt hier van alle brongegevens die, voor de AVG, gezien moeten worden als "persoonsgegevens" ook een gemaskeerde variant aangemaakt.

De gegevensstructuren in de bronzone zijn geoptimaliseerd voor gestandaardiseerde verwerking. Vandaar dat deze gegevens eenvoudiger gestructureerd worden in de **bronzone-ontkoppelviews**. Deze ontkoppelviews zorgen er ook voor dat het zichtbare deel van de bronzone-gegevens ook tijdens verwerking consistent blijft.

De gegevens in de **bronzone-ontkoppelviews** zijn, via **gegevensvensters** in de **bedrijfszone**, toegankelijk voor eindgebruikers. Wel worden deze bronzone-gegevens in de gegevensvensters beperkt tot alleen de gegevensset waar de eindgebruiker (of eigenlijk het proces waar die eindgebruiker het voor gebruikt) een rechtsgrond heeft, en waarvan het gebruik voldoet aan de AVG-principes m.b.t. proportionaliteit en subsidiariteit.

Deze gegevens in de **bronzone-ontkoppelviews** zijn te beschouwen als een 1:1 (maar wel vaak gemaskeerde) weergave van de gegevens in de bronnen. Noch in de **bronzone** zelf, noch in de **bronzone-ontkoppelviews** vindt namelijk data integratie plaats. Ook referentiële integriteit wordt in de bronzone niet afgedwongen. Binnen het DIM vallen dergelijke zaken, waar vereist, in de **integratiezone** en/of **bedrijfszone**.

Deze twee laatste zones maken de gegevens in het DIM namelijk op maat voor de afnemer:

- De **informatieproducten** in de **bedrijfszone** leveren de écht op maat gemaakte gegevens[33].
- Waar gegevens voor meerdere **informatieproducten** op dezelfde wijze bewerkt moeten worden (om redenen van efficiëntie of consistentie) gebeurt dat in de **integratie-zone**.

Voor de volledigheid:
Uiteindelijk vindt alle levering aan eindgebruikers plaats via **informatieproducten** in de **bedrijfszone**, ook als deze gegevens niet, of maar beperkt, hoeven te worden voorbewerkt, en dus direct kunnen worden betrokken uit **bronzone-ontkoppelviews** en/of **integratiezone**. In dergelijke gevallen ligt er namelijk altijd een **gegevensvenster** (in de **bedrijfszone**) tussen data

[33] Dat kunnen dus ook onbewerkte gegevens zijn: een gegevensvenster op (een deel van) de bronzone

en afnemer. Dit gegevensvenster is overigens een virtuele structuur. De data hoeft dus, behoudens bijzondere gevallen, niet in het gegevensvenster gerepliceerd te worden.

6.4 Versionering in de vier kern-zones

De **bronzone** is volledig geversioneerd en zal, voor die versionering altijd de administratieve tijdlijn van de bron volgen[34]. Alle overige door de bron aangeleverde historie-gegevens worden als functionele attributen gezien. De staging-laag (koppelvlak met de **ontsluitingszone**) volgt diezelfde benadering.

Versionering in de **integratiezone** en de **bedrijfszone** is niet altijd noodzakelijk; met name in de bedrijfszone voldoet de actuele situatie vaak. Waar versionering wel noodzakelijk is, zal deze in de regel ook weer gebaseerd zijn op een administratieve tijdlijn. Hier wordt alleen bij zeer specifieke afnemers-eisen van afgeweken.

6.5 Bronnen vs. bronmodules vs. leveringen

In de meeste gevallen levert een bron één bronlevering aan het DIM, in een vaste frequentie (dagelijks, wekelijks, maandelijks). Die leveringen bevatten in het algemeen gegevens van meerdere entiteiten binnen die bron.

De complete levering (dus alle entiteiten erin) wordt in de "bron-gerichte" zones van het DIM (**ontsluitingszone** en **bronzone**) in een eigen "silo" opgeslagen en verwerkt.

In sommige gevallen levert de bron voor dezelfde entiteiten twee leveringen: standaard een incrementele en in een lagere frequentie een volledige.
Hoe hier precies mee om te gaan wordt nog onderzocht[GL(4)].

In uitzonderlijke gevallen, waarbij in de bron sprake is van bronmodules met gescheiden functionaliteit en gegevens, zal dit ook in het DIM overgenomen worden in de vorm van aparte silo's per bronmodule (i.p.v. per bron). Dit is mogelijk omdat de verwerking van de gegevens uit bronmodules in de bronzone-data, vanwege de scheiding in de brongegevens, niet als één verversingsactiviteit hoeft te worden uitgevoerd.

6.6 Overzicht interactie met bronnen

Met de term "bron" kunnen verschillende zaken bedoeld worden:

- Een applicatie die gegevens aan het DIM levert
- De beheerder van die applicatie
- De (business) eigenaar van die applicatie
- De (business) eigenaar van de gegevens in die applicatie.

Interactie met een bron-**applicatie** is in overgrote deel van de gevallen eenrichtingsverkeer, de bron levert, via een of andere interface, gegevens aan het DIM.[35]
Alleen bij abonnementsgegevens is dit potentieel tweerichtingsverkeer: het DIM roept een service aan en de bron (de serviceleverancier) antwoordt.

Interactie met een bron-**beheerder** kent een "run" en een "change" smaak:

[34] Binnen UWV ook wel de "transactie-dimensie" genoemd. Zie https://digitalewerkplek.sharepoint.uwv.nl/documentcenter/Documenten/Gegevensdiensten/Beleidsdocument/Beleid%20Tijdsdimensies%20versie%201.4.pdf voor meer informatie.

[35] Het DIM geeft op dit moment geen succesmeldingen terug aan de bron-**applicaties**, aangezien alle leveringen via SIP-FT lopen. Indien nodig kan een dergelijk mechanisme in een later stadium aan de functionaliteit van de **ontsluitingszone** worden toegevoegd.

- **Run:**
  Succes- en foutmeldingen na het ontvangen en verwerken van door de beheerde bron-applicatie geleverde gegevens.
  Een bijzonder geval hiervan is de melding "niets ontvangen" als het DIM vergeefs op een levering uit een bronapplicatie heeft gewacht.
- **Change:**
  Verzoek om technische aanpassingen aan de levering. Dit verzoek kan uitgaan van DIM-beheer, of van de bron-beheerder.

Interactie met (business) **eigenaar** van een bron kent deze tweedeling ook:

- **Run:**
  Escalaties bij langdurige(r) leveringsproblemen.
- **Change:**
  Verzoek om functionele (of verregaande technische) aanpassingen aan de levering (of een nieuwe levering). Dit verzoek kan uitgaan van DIM-beheer, van de DIM-eigenaar[36] of van de bron-eigenaar.

En ook interactie met de **gegevens-eigenaar** van een bron kent deze tweedeling:

- **Run:**
  Escalaties bij datakwaliteitsproblemen (door de DIM-eigenaar, al dan niet namens de DIM-afnemers)
- **Change:**
  Verzoek om toestemming voor de opslag van brongegevens in het DIM, en het gebruik ervan door DIM-afnemers.

N.B. Alleen inrichting van de interactie met de bron-applicaties en de run-interactie met de bron-beheerders zijn in scope van het project DataFabriek. Alle overige interacties worden onder verantwoordelijkheid van de staande organisatie (Afdeling DWH en Gegevensdiensten Leveren) ingericht.
Het project maakt wel gebruik van bestaande "change"-processen van UWV bij het aanvragen van nieuwe en gewijzigde bronleveringen.

6.7 Overzicht interactie met afnemers

Met de term "afnemer" kunnen verschillende zaken bedoeld worden:

- Een UWV-medewerker die een informatieproduct van het DIM gebruikt
- Een applicatie waaraan het DIM gegevens levert (via een informatieproduct)
- De beheerder van die applicatie
- De (business) eigenaar van die applicatie

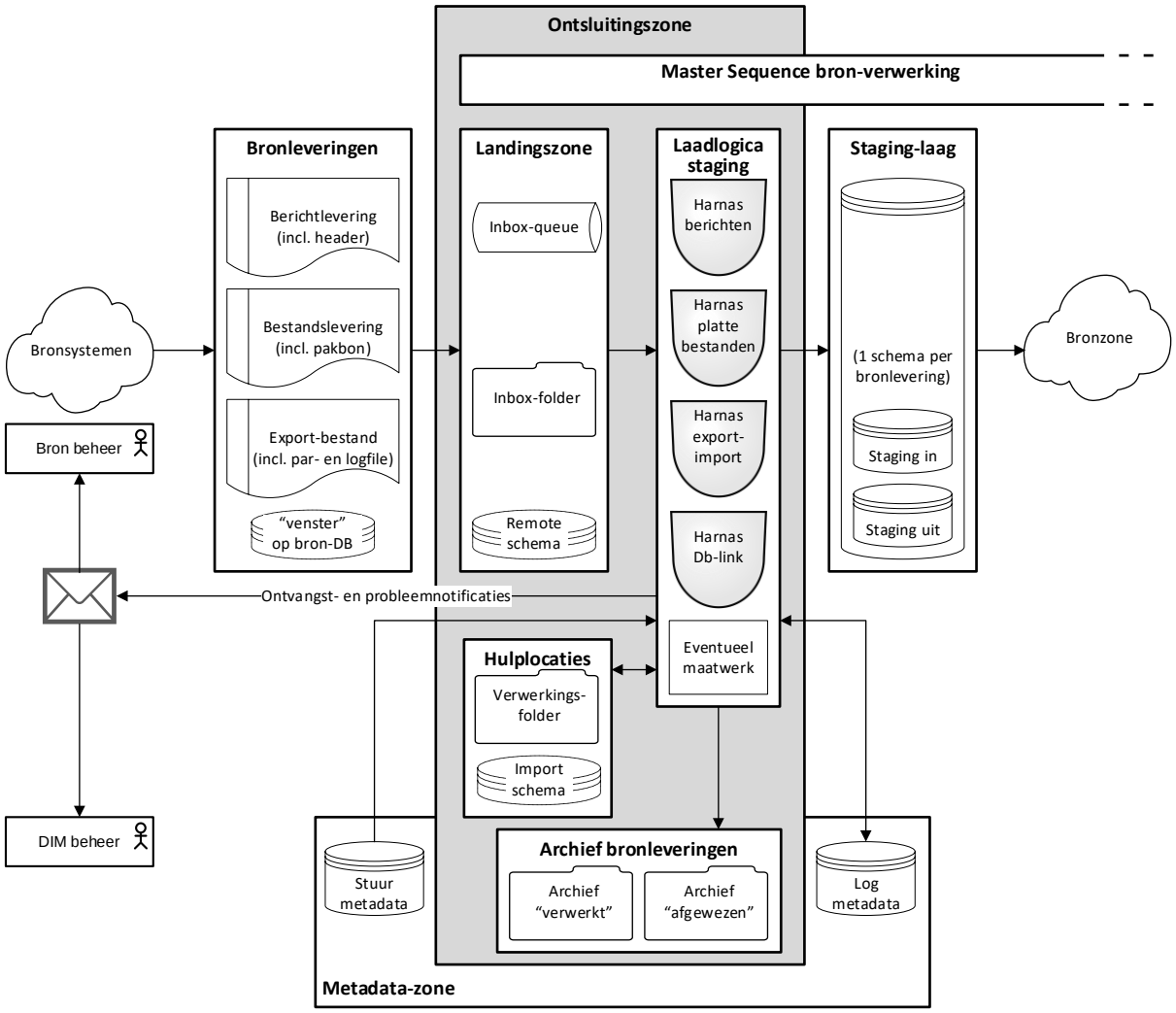
Inrichting van de levering van informatieproducten (aan mensen en machines) is in scope van het project DataFabriek. Alle overige interacties wordt ingericht door de staande organisatie.

Zodra daar meer duidelijk over is zal deze paragraaf worden aangepast.

[36] Opzetten van dit eigenaarschap is onderdeel van het "invlechten" van de afdeling DWH in de divisie Gegevensdiensten.

7 ONTSLUITINGSZONE EN STAGING-LAAG

7.1 Basisopzet



De Ontsluitingszone is de zone waarin de gegevensleveringen vanuit de bronnen, in welke technische vorm dan ook, binnenkomen in het DIM:

- Op **berichten** gebaseerde bronleveringen worden opgevangen in een inbox-queue;
- Op **bestanden** gebaseerde bronleveringen (bestandsleveringen en export/import) worden opgevangen in een folderstructuur welke o.a. de inbox en het archief bevat;
- Op **databaselinks** gebaseerde bronleveringen worden niet fysiek ontvangen in de ontsluitingszone; in plaats daarvan acteert het remote schema waar het DIM mee connecteert als "ontvangstpunt".

N.B. Deze bronleveringen zijn óf t.b.v. rapportage/analyse (meestal "brede bronontsluitingen", zie paragraaf 4.1 (Gegevens als basis – brede bronontsluiting), óf t.b.v. beveiliging, zie paragraaf 4.3 (Gegevens t.b.v. beveiliging).
Parameters voor informatieproducten (waaronder ook abonnementen, zie paragraaf 4.2, Gegevens als parameters) worden **niet** via de ontsluitingszone en de bronzone verwerkt, maar pas bij het aanmaken/verversen van het informatieproduct opgehaald (zie hoofdstuk 10, Bedrijfszone).

Alle ontvangstpunten gezamenlijk vormen de **landingszone**.
Alle bronleveringen, ongeacht hun techniek, komen "pushed" in die landingszone. Ze worden dus niet door het DIM opgehaald. De enige uitzondering is levering via een databaselink. Daar wordt de data opgehaald door het DIM, maar doet dat op een moment dat zulks, op aangeven van de bron, "veilig"[37] is. Ideaal is als de bron een "all clear"-signaal stuurt (als trigger-bestand naar de inbox-folder of als bericht naar de inbox-queue), maar lezen in een vast tijds-"window" is ook een optie, al dan niet gecombineerd met het sluiten van de database-link (door de bron) als lezen toch niet veilig is.

*De precieze inrichting van bronleveringen kan verschillen. Zie het document **Interface-standaarden (bron-DIM)** voor details.*

Binnen de ontvangstzone worden de bronleveringen door de "laadlogica naar de staging laag" (kortweg: **laadlogica staging**) "uitgepakt". Deze laadlogica wordt, waar mogelijk, uitgevoerd door **harnassen**, die op hun beurt weer gedreven worden door **stuur-metadata**, en hun verwerkingsresultaten (succes, ondervonden fouten, etc) opslaan in de **log-metadata**.

De uitgepakte (meta)data wordt, bij succesvolle verwerking, opgeslagen in het koppelvlak tussen de ontvangstzone en de bronzone: de **staging-laag**.

Deze **staging-laag** legt, als koppelvlak, een "knip" tussen het "uitpakken" van de bronleveringen en het verwerken daarvan richting de **bronzone**.

Dit maakt het mogelijk om ook een knip te leggen tussen het verwerken van de bronleveringen zelf (uitpakken, controle volgordelijkheid, etc) en het verwerken van de gegevens in die leveringen in de bronzone (versioneren, maskeren, etc), en zo de verwerking van die gegevens onafhankelijk uit te voeren van de wijze waarop ze geleverd zijn.

De **laadlogica bronzone** wordt opgestart door de **master sequence bron-verwerking**. Dit is een overkoepelende, door een scheduler opgestarte, aansturingscomponent die zowel de **laadlogica staging** (in deze zone) als de **laadlogica bronzone** (in de bronzone) aanstuurt.

Zie hoofdstuk 8 (Bronzone en ontkoppelviews~~Bronzone en ontkoppelviews~~) voor meer informatie over de master sequence en de verwerking richting bronzone.

Voorafgaand aan de verwerking van elke nieuwe bronlevering worden de oude gegevens van die bronlevering uit de **staging-laag** verwijderd.

Na uitpakken van de bronlevering in de **staging laag** wordt de levering zelf (of, bij database-links, een dump daarvan) opslagen in het **archief bronleveringen**.

7.2 Landingszone, hulplocaties en archief

Voor elke bron(module) bevat de ontsluitingszone een folderstructuur, met daarin de volgende folders:

- **inbox-folder**
  In deze map levert de bron zijn pakbon- en databestanden.
  Voor bronleveringen via een database-link wordt deze folder alleen gebruikt als de bron een trigger-bestand levert.
- **verwerkings-folder**
  In deze map staan de bestanden op het moment dat ze worden verwerkt door het DIM.
  Voor bronleveringen via een database-link wordt deze folder niet gebruikt.

[37] Zowel qua gegevensconsistentie als qua bronsysteembelasting

- **archief-folders** voor **verwerkte** en **afgewezen** leveringen
  Als het verwerkingsproces succesvol is afgerond worden de bestanden naar de archief-folder voor verwerkte leveringen verplaatst. Is de levering afgewezen, of afgebroken tijdens de verwerking, dan worden de bestanden verplaatst naar de archie-folder voor afgewezen leveringen. Als er tijdens het verwerkingsproces een reject-bestand is aangemaakt[38], dan staat ook dit reject-bestand in deze archief-folder.
  Bij bronleveringen via een database-link levert de bron geen bestanden.
  Daarom archiveert het DIM in dat geval Oracle-exports van de tijdens de verwerking uit de bron opgehaalde, en vervolgens in de staging-laag gekopieerde, gegevens.

Naast bovenstaande folder-structuur bevat de ontsluitingszone ook twee RDBMS-bouwstenen:
- Voor bronleveringen via een database-link fungeert het remote schema waar het DIM mee connecteert als landingszone (i.p.v. de inbox-folder).
- Voor bronleveringen via exportbestanden bestaat, naast de verwerkings-folder, ook een **importschema**, waarin de geleverde dump-bestanden worden geïmporteerd.

N.B. In sommige gevallen levert de bron voor dezelfde entiteiten twee leveringen: bijvoorbeeld standaard een incrementele en in een lagere frequentie een volledige. Deze leveringen worden in aparte silo's[39] verwerkt (ze komen pas samen in de bronzone), en hebben dus ook elk een eigen folderstructuur.
Er loopt nog een onderzoek [GL(5]*om te kijken of deze twee leveringen niet toch in één folderstructuur (en één silo) moeten worden verwerkt. Dit vergroot de complexiteit van de verwerking binnen de ontsluitingszone, maar verlaagt de complexiteit van de verwerking in de bronzone, en van de routering van de levering (door bron en SIP-FT).*

7.3 Staging-laag

De **staging-laag** is een **technische** data-laag, alleen bedoeld als tussenstap in de totale verwerking van bronlevering naar bronzone. De staging-laag is dus niet bedoeld (of geschikt) als bron voor informatieproducten, en is dan ook alleen toegankelijk voor (de technische user accounts gebruikt door) ETL-programmatuur.[40]

De **staging-laag** legt, als koppelvlak, een "knip" tussen het "uitpakken" van de bronleveringen in de **ontsluitingszone** en het verwerken daarvan richting de **bronzone**.

Dit maakt het mogelijk om ook een knip te leggen tussen het verwerken van de bronleveringen zelf (uitpakken, controle volgordelijkheid, etc) en het verwerken van de gegevens in die leveringen in de bronzone (versioneren, maskeren, etc), en zo de verwerking van die gegevens onafhankelijk uit te voeren van de wijze waarop ze geleverd zijn.

Bij het verwerken van een bronlevering zullen alle verwerkingen die onafhankelijk van de leveringstechniek uitgevoerd kunnen worden dus ook pas uitgevoerd worden als de gegevens, in de staging-laag, ook werkelijk onafhankelijk zijn gemaakt van die leveringstechniek.
Veel functionaliteit hoeft dus maar op één plek (in de bronzone) onderhouden te worden.

Bijkomend voordeel is dat zelfs in het geval dat er, in de ontsluitingszone, maatwerk-ETL gemaakt moet worden voor het verwerken van een bronlevering dit beperkt blijft tot het op een correcte wijze vullen van de staging-laag; verwerking vanaf dat punt (in de bronzone) kan vervolgens weer op de standaard-wijze worden afgehandeld.

---

[38] Zie paragraaf 7.4 (Laadlogica staging)

[39] Zie paragraaf 6.5 (Bronnen vs. bronmodules vs. leveringen)

[40] En, in geval van calamiteiten, voor DIM Beheer (t.b.v. foutenonderzoek)

Zie hoofdstuk 8 (Bronzone en ontkoppelviews~~Bronzone en ontkoppelviews~~) voor meer informatie over de verwerking van de gegevens in de staging-laag door (het harnas in) de bronzone.

De **staging-laag** bevat voor elke bron(module) een eigen schema met tabellen. Dat schema bevat alleen de geleverde **data**; de geleverde **metadata** wordt, tijdens de verwerking, opgenomen in de **log-metadata**.

De tabellen in de staging-laag bevatten alleen tijdelijke gegevens, gebruikt bij de verwerking van de bronlevering. Voorafgaand aan de verwerking van elke nieuwe bronlevering worden de oude gegevens van die bronlevering uit de staging-laag verwijderd.

7.3.1 Structuur van de staging-laag

De **staging-laag** bevat voor elke bron(module) een eigen schema.

De tabellen in dat schema vallen in twee blokken uiteen:

- **Staging in -** Tabellen met de onbewerkte aangeleverde data
  Deze tabellen zijn qua structuur gelijk aan wat er zou moeten worden aangeleverd (conform de RLO en vervolgens vastgelegd in de stuur-metadata). Hieraan zijn geen technische velden toegevoegd.
- **Staging uit -** Tabellen waarin de aangeleverde data is omgezet naar de hub/satelliet-structuur van de bronzone-data[41] (ook wel bekend als de "technische staging tabellen").

De **laadlogica staging** (zie hieronder) laadt **staging in**; het omzetten van de gegevens van **staging in** naar **staging uit** wordt uitgevoerd binnen de **laadlogica bronzone**.[42]

7.4 Laadlogica staging

Voor elke bronlevering (en ongeacht de leveringstechniek) doorloopt de **laadlogica staging** in grote lijnen dezelfde stappen:

1 Start

De **master sequence bronverwerking**[43] start de ETL-job waarin de laadlogica wordt uitgevoerd (meestal een harnas, soms maatwerk), en geeft daaraan de benodigde inputparameters (Run ID, Bronlevering ID, Bronlevering Naam, Bron Naam) mee.

Verder wordt er ook een omgevingsafhankelijke default parameter-set meegegeven. De inhoud van deze parameter-sets is gebaseerd op de omgevingsvariabelen die op DIM project-niveau in DataStage gedefinieerd zijn (zie paragraaf 8.6.1).

De master sequence wordt standaard gescheduled (op basis van de leverafspraken met de bron). Bij reruns t.b.v. herlevering wordt de master sequence handmatig gestart (door DIM Beheer).

Haal, m.b.v. de door de master sequence meegegeven parameters, de voor de verwerking van de bronlevering benodigde stuur-metadata op.

Maak een record aan in de log-metadata met als status "lopend".

[41] Zie paragraaf 8.2 (Bronzone-data) voor meer info.

[42] Deze logica wordt beschreven in paragraaf 8.3 (Laadlogica bronzone).

[43] Zie paragraaf 8.6 (Master Sequence bronverwerking~~Het hele laadproces in één hand – De Master Sequence~~) voor meer info.

2 Initialisatie

Bepaal, op basis van de DataStage omgevingsvariabelen, de omgevingsafhankelijke[44] delen van paden en bestandsfolders.

3 Detecteer beschikbaarheid levering

a) Voor levering middels **platte bestanden** en levering middels **export-bestanden** gebeurt dit o.b.v. "polling" op de aanwezigheid van het pakbon-bestand resp. de par-file in de **inbox-folder** voor de betreffende levering (deze bestanden worden als laatste onderdeel van de bronlevering aangeleverd en kunnen daarom als trigger voor de verwerking van de bronlevering dienen)

De naam en locatie van de inbox-folder kan worden afgeleid uit de combinatie van omgevingsvariabelen en stuur-metadata, het te gebruiken "filemask"[45] staat in de stuur-metadata.

b) Voor levering middels een **database-link** dient de bron een "all clear" signaal te leveren. Het precieze proces van deze signalering kan per bron verschillen.[46]
In de stuur-metadata staat aangegeven welke checkmethode moet worden gebruikt om te valideren dat de database-link klaar is voor verwerking.
Voor leveringen middels **berichten** gaat dit gebeuren[47] door te kijken of er berichten in de **inbox-queue** staan. Als er geen berichten zijn is dit overigens i.h.a. géén foutsituatie; blijkbaar waren er in de achterliggende periode geen gegevenswijzigingen in de bron.
In de stuur-metadata staat aangegeven welke server/queue-combinatie gebruikt wordt als inbox-queue. Het tijdsvenster waarbinnen een levering beschikbaar zou moeten zijn staat in de stuur-metadata. Als een levering niet binnen dit tijdsvenster verschijnt zal de levering beschouwd worden als niet beschikbaar (en kan deze stap dus worden afgebroken).

Sla de uitkomsten van deze stap, en eventuele overige relevante procesinformatie, op in de **log-metadata**.
[indien levering niet beschikbaar: door naar stap X1]

4 Zet de levering klaar voor verwerking

a. Voor levering middels **platte bestanden** gebeurt dit door de bestanden (platte bestanden + pakbon) van de **inbox-folder** naar de **verwerkingsfolder** te verplaatsen.
b. Voor levering middels **export-bestanden** gebeurt dit door de bestanden (dumpbestanden + par- en logfile) eerst van de **inbox-folder** naar de **verwerkingsfolder** te verplaatsen, de dumpfiles vervolgens in het (daarvoor geschoonde) **importschema** te importeren.
c. Voor levering middels een **database-link** wordt deze stap overgeslagen.

Sla de uitkomsten van deze stap, en eventuele overige relevante procesinformatie, op in de **log-metadata**.

---

[44] Met "omgevingsafhankelijk" wordt bedoeld: afhankelijk van de OTAP-omgeving waarin het proces draait.
[45] Een filemask is een reguliere expressie (regexp) welke gebruikt kan worden om bestandsnamen door middel van een patroon te herkennen
[46] Ideaal is als de bron een "all clear"-boodschap stuurt (als trigger-bestand naar de inbox-folder of als bericht naar de inbox-queue), maar lezen in een vast tijds-"window" is ook een optie, al dan niet gecombineerd met het sluiten van de database-link (door de bron) als lezen toch niet veilig is.
[47] Er zijn in fase 4/5 geen op berichten gebaseerde bronleveringen

5 Sla pakbon-metadata op

Sla de geleverde tracking- en controle-metadata op in de **log-metadata**.

a. Bij leveringen middels **platte bestanden** komt deze metadata uit het pakbonbestand
b. Bij leveringen middels **export-bestanden** komt deze metadata uit de par- en logfiles.
c. Bij leveringen middels een **database-link** komt deze metadata deels uit de brondatabase-metadata (de "systables"), en deels uit, via het remote-schema ontsloten en speciaal voor deze levering door de bron aangemaakte, "pakbon-tabellen".

Hier wordt op basis van o.a. de peil- of einddatum ook de procesdatum bepaald, benodigd voor de verdere verwerking. De exacte methode verschilt per leveringswijze.

*Sla de uitkomsten van deze stap, en eventuele overige relevante procesinformatie, op in de **log-metadata**.*

De pakbon, par- en log-file worden in een blob formaat opgeslagen zodat ze zonder wijzigingen beschikbaar blijven. Deze en eventuele overige relevante procesinformatie, worden in de log-metadata opgeslagen.

[indien de pakbon, parfile/logfile of metadata-view technisch niet verwerkt kan worden: door naar stap X1]

6 Controleer de pakbon-metadata

a. Voldoet de levering qua geleverde objecten/velden aan wat er, op basis van de RLO, is vastgelegd in de stuur-metadata?
   Het betreft hier alleen controles die noodzakelijk zijn voor een betrouwbare verdere verwerking richting bronzone, dus, bijvoorbeeld, geen controles of een veld voldoet aan het in de RLO ervoor gedefinieerde waardebereik.
b. Sluit de levering qua peildatum/periode aan op de vorige levering?
   Controle op basis van stuur-metadata en log-metadata (tracking-metadata van deze en de vorige levering).
c. Eventuele extra controles (niet standaard ingericht).
   Voorbeeld:
   - of uit de parfile blijkt dat een database-export, conform de interface-standaarden, als consistente set is aangemaakt, door gebruik van de FLASHBACK-TIME-optie van Datapump.of de primaire sleutel (zoals, op basis van de RLO, gedefinieerd in de stuur-metadata) inderdaad uniek is

   Deze controles zullen, als optionele controles, aan harnassen en stuur-metadata worden toegevoegd zodra ze vereist blijken. De harnassen zijn zo ingericht dat het toevoegen van controles relatief eenvoudig is, en alleen impact heeft op deze stap binnen het harnas.

Sla de uitkomsten van deze stap, en eventuele overige relevante procesinformatie, op in de **log-metadata**.

[indien levering niet correct: door naar stap X1]

7 Schoon de staging-laag

Truncate de tabellen voor deze levering, zodat de verwerking met een "schone lei" kan beginnen.
Als uit de (voor de datum/tijd van de bronlevering geldige versie van de) stuur-metadata blijkt dat de bronlevering een nieuwe structuur heeft, vervang dan de bestaande tabellen in **staging in** door nieuwe (met de nieuwe structuur).

Creëer, indien noodzakelijk, de DataStage schema files die nodig zijn om platte bestanden te kunnen laden.

8 Laad de staging-laag

Laad de gegevens uit de platte bestanden, het importschema of het remote schema in de staging laag (in **staging in**).

Hierbij vindt er op record-niveau automatisch een check plaats of de te laden gegevens overeenkomen met de in de stuur-metadata gedefinieerde verwachte structuur.

Maak, t.b.v. fout-analyse, in de **verwerkingsfolder** een **reject-bestand** aan met de rijen die (eventueel) niet geladen kunnen worden.

Sla de uitkomsten van deze stap, en eventuele overige relevante procesinformatie, op in de **log-metadata**.

[indien (deels) niet geladen data: door naar stap X1; alleen volledig laadbare leveringen worden verder verwerkt]

9 Controleer kentallen

Controleer of de geleverde datalevering consistent is met de kengetallen (bv. aantal rijen) verkregen uit de pakbon, par/log file of "pakbon" query op remote schema.

Eventuele extra controles (niet standaard ingericht, bijvoorbeeld:

- of de primaire sleutel (zoals, op basis van de RLO, gedefinieerd in de stuur-metadata) inderdaad uniek is (en altijd gevuld);
- of het aantal geleverde rijen niet significant afwijkt van de vorige levering (voor bronnen die af en toe onvolledig blijken te leveren).

Deze stap moet er, samen met stap 6, voor zorgen dat alleen gegevens uit betrouwbaar gebleken bronleveringen worden verwerkt in de bronzone. Dit om de kans op het uit moeten voeren van "undo's" te minimaliseren.

Sla de uitkomsten van deze stap, en eventuele overige relevante procesinformatie, op in de **log-metadata**.

[indien inconsistente levering: door naar stap X1]

X1 Archiveer de tot hier verwerkte bronlevering
Doe dit voor succesvol t/m de staging-laag verwerkte leveringen in de archief-folder voor verwerkte leveringen, en voor niet verwerkbare leveringen in de archief-folder voor afgewezen leveringen. Archiveer in dit laatste geval ook de reject-bestanden naar de archief-folder voor afgewezen leveringen.

a. Voor levering middels **platte bestanden** en levering middels **export-bestanden** gebeurt dit door de bestanden (platte bestanden + pakbon, resp. dumpbestanden + par- en logfile) van de **verwerkingsfolder** naar de juiste **archief-folder** te verplaatsen.
b. Voor levering middels **database-links** gebeurt dit door de relevante tabellen in de staginglaag te exporteren, en de resulterende dumpbestanden in de juiste **archief-folder** te plaatsen.

X2 Verzend notificaties
Meld succes of issues via e-mail aan de relevante partijen.
Welke meldingen naar welke partijen moeten, en welke emailadressen voor deze partijen gebruikt moeten worden, staat in de stuur-metadata.

Op dit moment krijgt alleen DIM-beheer notificaties. In fase 5 worden de notificatie-vereisten verder uitgewerkt, en wordt deze stap (en de bijbehorende stuur-metadata) aangepast om die vereisten te ondersteunen. Onderdeel van deze uitwerking zal ook zijn hoe met reject-bestanden (met daarin potentieel persoonsgegevens) moet worden omgegaan.

X3 Einde job
Vervang de log-markering "lopend" door "klaar" of "afgebroken", afhankelijk van het resultaat van de verwerking tot op dit punt.
[einde ETL-job]

De master sequence bepaalt nu of de verdere verwerking naar de bronzone gestart wordt, of dat de master sequence stopt.

N.B. De **laadlogica staging** bevat alle controles die nodig zijn om te voorkomen dat er problemen ontstaan bij de verwerking van de staging-laag in de **bronzone**.
Als een bronlevering op één aspect (of voor één tabel) niet voldoet, dan zal de hele bronlevering worden afgewezen, en zal er dus ook géén vervolgverwerking in de bronzone plaatsvinden.

Verwerking van bronleveringen o.b.v. berichten is hierboven nog niet beschreven, aangezien dit in de projectfasen 4 en 5 nog niet relevant is.

7.5 Harnassen

In het interfacestandaarden document worden de verschillende methoden om bronleveringen te doen beschreven.

De standaarden die in dit document staan beschreven maken het mogelijk om ook de verwerking van deze leveringen binnen het DIM te standaardiseren.
Dit gebeurt middels generieke, en metadata-gedreven, ETL-jobs: de **bronleveringsharnassen**.

Voor de verwerkingen binnen de **ontsluitingslaag** (van **bronlevering** naar **staging-laag**) worden vier harnassen gebruikt, voor elke leveringstechniek één:

- harnas platte bestanden
- harnas export-import

- harnas databaselink
- harnas berichten

De door deze harnassen uit te voeren stappen zijn in grote lijnen voor alle harnassen gelijk (zie vorige paragraaf). De harnassen maken daarom, waar mogelijk, gebruik van één keer gebouwde , en binnen meerdere harnassen bruikbare, ETL-bouwstenen.

N.B. Bij het databaselink-harnas is er geen sprake van fysieke bestanden die aangeleverd worden. Er worden dus ook geen bestanden in de staging-laag ingelezen. In plaats daarvan kopieert het harnas de inhoud van tabellen in de brondatabase (via het remote schema waarop de database-link rechten heeft) direct naar de staging-laag.

N.B. Het berichten-harnas wordt voorlopig niet ontwikkeld daar er momenteel geen bronnen voorzien zijn die berichten gaan sturen naar het DIM. Er is in de stuur-metadata wel voorzien dat dit mogelijk moet zijn.

Als een bronlevering dusdanig afwijkt van de standaard dat deze niet door een harnas verwerkt kan worden (of het harnas daarvoor onacceptabel zou compliceren), dan wordt voor die levering een **maatwerk ETL-job** ontwikkeld. Startpunt voor dat maatwerk is een kopie van het harnas voor de betreffende leveringstechniek. Ook een maatwerk-job is (zoveel mogelijk) metadata-gedreven en maakt, waar mogelijk, gebruik van de hierboven genoemde bouwstenen.

Zie hoofdstuk 15 (Generieke oplossingen~~Generieke oplossingen~~) voor meer informatie over de algemene opzet van ETL-harnassen en ETL-bouwstenen.

7.6 Technische aspecten

7.6.1 Toegangsrechten

Alle folders en database-objecten in de ontsluitingszone en de staging-laag zijn (op A en P) alleen toegankelijk voor geautomatiseerde processen.

Hierop zijn twee uitzonderingen:

- DBA's kunnen folders (en in de toekomst potentieel ook queues) aanmaken, t.b.v. het aansluiten van nieuwe bronnen, en database-objecten wijzigen t.b.v. het aansluiten van nieuwe bronnen of het wijzigen van bestaande leveringen.
  Idealiter loopt beide overigens via een door een release-tool uitgevoerd script, zodat deze toegang niet noodzakelijk is
- DBA's, beheerders en/of ontwikkelaars kunnen, in geval van een calamiteit, tijdelijk toegang krijgen tot folders en database-objecten t.b.v. onderzoek. Deze tijdelijke toegang verloopt via een "red envelope"-procedure.[48]

[48] Zie hoofdstuk 18 (Informatiebeveiliging en –beheer) voor meer informatie over deze procedure.

7.7 Impact ontwerpuitgangspunten

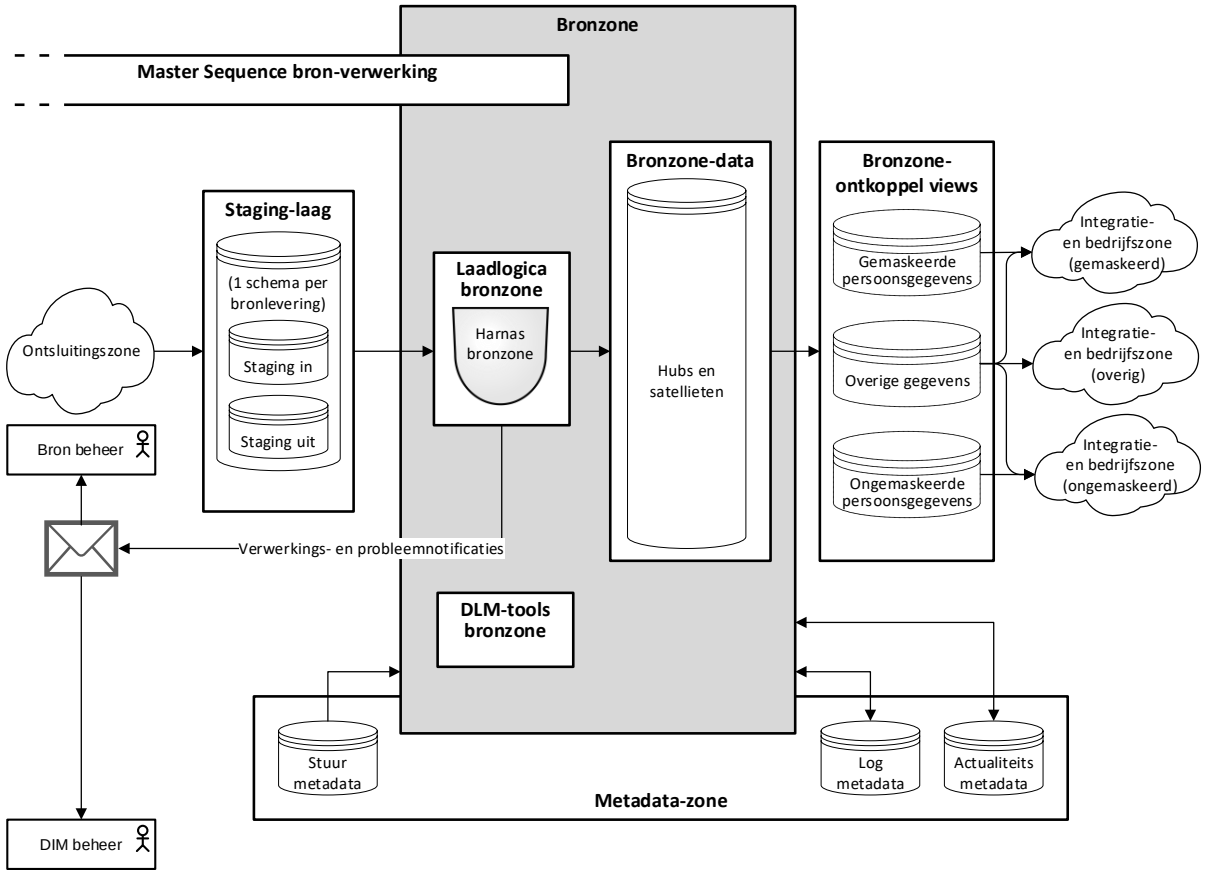
De ontwerpuitgangspunten zijn meegenomen in (dit deel van) het conceptueel ontwerp. Daarnaast moeten ze meegenomen worden in de detaillering en de realisatie van dat ontwerp.

Ontwerpuitgangspunt	**Impact**
1. Bewezen concepten	Opzet met staging-laag tussen inbox en bronzone is een "market best practice". Metadata-gedreven harnassen zijn een bewezen concept. De ervaring heeft overigens wel geleerd dat ze bij DataStage-gebaseerde data warehouses weinig worden toegepast
2. Geen realtime ambitie	Alle verwerking is batch-gedreven
3. Maximale ontkoppeling	Duidelijke scheiding door middel van staging-laag tussen de ontsluitingszone en de bronzone.
4. Specifiek waar mogelijk, gemeenschappelijk waar noodzakelijk of waardevol	Maximale inzet van generieke bouwstenen en harnassen, maar binnen die structuur ook mogelijkheid tot maatwerk.
5. Kortste keten	*(hier nog niet van toepassing)*
6. Compliant	Toegang tot folders en database (staging-laag, import-schema) alleen voor ETL-processen, tenzij expliciet noodzakelijk bij het afhandelen van calamiteiten.
7. Eenvoudig	Modulair opgebouwde harnassen, elk zo eenvoudig mogelijk gehouden door elke leveringstechniek door een eigen harnas te laten afhandelen.
8. Gebruiksvriendelijk	*(hier nog niet van toepassing)*
9. Gedefinieerd	Geen businesslogica in de ontsluitingszone; alle gegevens in de staging-laag zijn dus direct herleidbaar tot de bron. Daarnaast: alles-of-niets laden; daardoor geen "verborgen" inconsistenties met de bron.
10. Zoveel mogelijk RDBMS-onafhankelijk	Zoveel mogelijk gebruik maken van DataStage. SQL als tweede keus, geen PL/SQL of andere Oracle-specifieke functies (tenzij gegenereerd

	door DataStage). Dit bleek niet helemaal mogelijk te zijn. In de staging oplossing wordt namelijk gebruik gemaakt van een klein aantal Oracle stored procedures: • voor het truncaten van tabellen • voor het creëren van staging tabellen De procedures zijn noodzakelijk omdat we tijdens executie geen gebruik kunnen maken van de database schema's waarin de staging tabellen zijn gedefinieerd, en het onwenselijk is om systeembrede autorisaties uit te delen die de gewenste toegang tot deze schema's mogelijk zou maken (DROP ANY TABLE/CREATE ANY TABLE privileges). N.B. Export/import bronleveringen (datapump) zijn niet compliant met dit principe, maar worden om pragmatische redenen (minimaliseren impact bij de bronnen) toch ondersteund. De resulterende database-afhankelijkheid binnen de DIM-verwerking is overigens beperkt; deze loopt slechts tot aan het importschema.
11. Lineage mag niet gebroken worden	*(hier nu nog niet van toepassing)*
12. Minimale, en beheerbare, toolset	Oplossingen primair gebaseerd op DataStage, daarnaast bij de afdeling DWH bekende tools, zoals Oracle Datapump

8 BRONZONE EN ONTKOPPELVIEWS

8.1 Basisopzet



De **bronzone** is het hart van het DIM. Hier wordt de administratieve werkelijkheid van de DIM-bronnen opgeslagen en, waar nodig, voorzien van extra historie (in lengte of detail).
Daarnaast wordt hier van alle brongegevens die, voor de AVG, gezien moeten worden als "persoonsgegevens" ook een gemaskeerde variant aangemaakt.

In de **bronzone** vindt géén data integratie plaats, en wordt geen referentiële integriteit (of andere kwaliteitseis) afgedwongen. Binnen het DIM vallen dergelijke zaken, waar vereist, in de **integratiezone** en/of **bedrijfszone**.

De gegevens in de staging-laag worden door de **laadlogica bronzone** verwerkt in de **bronzone-data**.
Deze laadlogica wordt, waar mogelijk, uitgevoerd door het **harnas bronzone**, die op zijn beurt weer gedreven worden door **stuur-metadata**. De verwerkingsresultaten (succes, ondervonden fouten, etc) komen in de **log-metadata**. Na verwerking van een bronlevering wordt ook de **actualiteits-metadata** bijgewerkt, zodat (m.n. voor afnemende ETL-processen) eenvoudig te achterhalen is tot op welk punt in de tijd de gegevens in de bronzone zijn bijgewerkt.

In de **laadlogica staging** (in de **ontsluitingszone**), wordt elke leveringstechniek verwerkt door een eigen harnas. De **laadlogica bronzone** is echter zo ingericht dat alle verwerking van data van de **staging-laag** naar de **bronzone-data** (ongeacht de leveringsvorm) door één enkel harnas kan worden uitgevoerd.

De verwerking van brongegevens door de **laadlogica bronzone** wordt zo agile mogelijk gemaakt door, onder andere, gebruik te maken van Data Warehouse Automation (DWA) concepten; de ETL-harnassen.

De **laadlogica bronzone** wordt, net als de **laadlogica staging**, opgestart door de **master sequence bron-verwerking**..

Het Data Lifecycle Management van de bronzone-data wordt deels uitgevoerd door de **laadlogica bronzone**, en deels door de **DLM-tools bronzone**.

De **bronzone-data** bevat zowel gemaskeerde als ongemaskeerde gegevens. Deze gegevens zijn, anders dan in de **integratiezone** en **bedrijfszone**, niet volledig van elkaar gescheiden, maar gezamenlijk opgeslagen in, deels op DataVault2 gebaseerde, hub/satelliet-structuren. Dit om, bij de verwerking (door de **laadlogica bronzone**) van de (altijd ongemaskeerde) gegevens uit de **staging-laag**, het versioneren van beide typen **bronzone-data** zo gecontroleerd en efficiënt mogelijk te laten verlopen.

Het **koppelvlak** tussen de **bronzone** en de achterliggende zones (**integratiezone** en **bedrijfszone**) wordt gevormd door de **bronzone-ontkoppelviews**.
Deze ontkoppelviews dienen vier doelen:

1. Ze maken de technische structuur van de bronzone-data onzichtbaar voor de achterliggende zones; die zien géén hub/satelliet-structuren, maar "herkenbare", al dan niet geversioneerde, brondata.
2. Ze splitsen gemaskeerde en ongemaskeerde data; de achterliggende zones (en de afnemers daarvan) zien óf gemaskeerde óf ongemaskeerde persoonsgegevens (en in sommige gevallen geen van beide).
3. Ze verbergen de "zachte verwijderingen" voor het achterland; in de ontkoppelviews worden "zacht verwijderde" gegevens namelijk niet getoond. Doordat de laadlogica voor de **integratiezone en/of bedrijfszone** zich baseert op de ontkoppelviews zullen "zachte verwijderingen" dus als "harde verwijderingen" doorwerken naar de **integratiezone** en de **bedrijfszone.**
4. Ze ontkoppelen het verversen van de bronzone-data van het gebruiken ervan; achterliggende zones (en de afnemers daarvan) blijven tijdens verversen de "oude" situatie (voor verversen) zien. Na verversen "flippen" alle ontkoppelviews voor een bronlevering in één keer, gezamenlijk, over naar de nieuwe, volledig ververste, situatie.

N.B. Met name vanwege punt 2 loopt toegang tot de bronzone-data **altijd** via de ontkoppelviews (ook voor de laadprocessen van integratiezone en bedrijfszone), en toegang buiten de ontkoppelviews om zal als een (potentieel) datalek moeten worden beschouwd.
De enige uitzondering daarop is toegang door DIM-beheer (t.b.v. het analyseren van foutsituaties)

N.B. Het DWH3-equivalent van de bronzone-ontkoppelviews is de Materialized View Laag (MVL). Grote verschil tussen de MVL en de (ook gematerialiseerde) ontkoppel-views is echter dat die laatste:

- de gegevens uit de bronzone volledig en verplicht afschermen van de rest van het DIM;
- wel gegevens uit de bronzone repliceren, maar per definitie géén business logica bevatten.

De ontkoppelviews zijn er dus, zoals de naam al zegt, vooral om de bronzone te ontkoppelen van de rest van het DIM (je zou het zelfs een veredelde cache kunnen noemen), en vormen géén aparte functionele laag binnen het DIM.

8.2 Bronzone-data

De modellering van de bronzone-data gebruikt een DIM-specifieke opzet, gericht op maximale scheiding van gemaskeerde en ongemaskeerde gegevens, maar wel voldoende gestandaardiseerd om generieke laadprocessen mogelijk te maken. In die modellering worden concepten uit DataVault2 gebruikt. De data mag echter niet gezien worden als "gemodelleerd conform DataVault2". Dit omdat de structuur ook duidelijk (en bewust) van DataVault2 afwijkt.

De verwerking van brongegevens (in de **staging-laag**) door de **bronzone** wordt zo agile mogelijk gemaakt door, onder andere, gebruik te maken van Data Warehouse Automation (DWA) concepten zoals (stuur)metadata-gedreven ETL-harnassen.

Om metadata-gedreven harnassen mogelijk te maken, conformeert de tabelstructuur waarin een bron-entiteit wordt opgeslagen zich voor elke entiteit aan hetzelfde standaardpatroon. Om aan AVG en Archiefwet te kunnen voldoen is dit patroon zo gestructureerd dat:

- Identificerende, gemaskeerd identificerende, en niet-identificerende gegevens in aparte tabellen zijn opgeslagen;
- Identificerende gegevens kunnen worden verwijderd zonder dat dit ten koste gaat van samenhang of detailniveau van de gemaskeerde en niet-identificerende gegevens;
- Toegang tot identificerende gegevens eenvoudig kan worden afgeschermd.

De bronzone volgt de administratieve werkelijkheid van de bron, en kan worden beschouwd als een 1:1 (maar wel vaak gemaskeerde) weergave van de (historie van) gegevens in die bron.

Ook als de geleverde gegevens van lage kwaliteit zijn worden ze verwerkt in de bronzone; die lage kwaliteit is immers de "administratieve werkelijkheid".

N.B. In een beperkt aantal gevallen zijn afgeleide gegevens al in de bronzone noodzakelijk.
Deze afgeleide velden zijn per definitie een aanvulling op de 1:1 weergave; de voor de afleiding gebruikte input is altijd ook onderdeel van de bronzone-data.
Zie verder paragraaf 8.3.4 (Aanmaken van afgeleide velden).

8.2.1 Identificerende data en niet-identificerende data

In de bronzone worden identificerende gegevens anders verwerkt en opgeslagen dan niet-identificerende gegevens.

Gegevens zijn identificerend als ze direct of in combinatie met andere gegevens ertoe leiden dat de persoon waarop de gegevens betrekking hebben geïdentificeerd kan worden. De rest van de gegevens is niet-identificerend.

Of een gegeven identificerend is wordt afgeleid uit de RLO[49] en staat, per veld/kolom, in de **stuur-metadata**.

1.1.1 [49] Zie Interface-standaarden en paragraaf 12.2.1~~12.3~~(Metadata-componenten o.b.v. documenten

Een aantal metadata-componenten zijn eigenlijk verzamelingen van documenten (meestal conform een vast sjabloon), opgeslagen in een vaste folder-structuur. Ze worden vooral gebruikt om afspraken tussen verschillende partijen en/of teams vast te leggen.

1.1.1.1 GIA en GLA

De GegevensInwinningsAfspraak (GIA) is het contract waarin de afspraken rondom een levering van gegevens door een bron **aan** het DIM zijn vastgelegd.

N.B. De term "identificerend" moet hierbij vanuit het perspectief van de burger gezien worden. Identificerende gegevens betreffende UWV-medewerkers (in hun rol als medewerker, dus niet in hun rol als burger) worden in de bronzone niet gemaskeerd.

Voor gegevens die geen betrekking hebben op burgers zijn alle gegevens dus, in de bronzone, "niet-identificerend".

In sommige gevallen is een veld "soms identificerend", bijvoorbeeld als het, in de bron, als een "generiek" veld is gedefinieerd, en de betekenis ervan (en dus ook of het identificerend is of niet) vastgelegd is in een ander veld. Dergelijke velden worden in de bronzone beschouwd en verwerkt als identificerend".[50]

8.2.2 Datamodel

Het datamodel van de bronzone is "op maat ontworpen" om gemaskeerde en ongemaskeerde gegevens met minimale gegevensduplicatie toch maximaal te scheiden. In het model worden (varianten op) een aantal uit DataVault2 ontleende concepten en begrippen toegepast, zoals hub/satelliet-structuren en hashkeys.
De bronzone is echter uitdrukkelijk géén DataVault2-implementatie; het datamodel is primair gebaseerd op de behoeften van het DIM, en niet op de Datavault2-theorie.
De volgende paragrafen beschrijven het datamodel, inclusief overeenkomsten en verschillen met DataVault2, in meer detail.

8.2.2.1 Hubs en satellieten

Een tabel in de **staging-laag** wordt in de **bronzone-data** altijd in het volgende standaardpatroon opgeslagen:

- Hub (H)
- Satelliet met ongemaskeerde identificerende gegevens (ID)
- Satelliet met gemaskeerde identificerende gegevens (ID-M)
- Satelliet met niet-identificerende gegevens (N-ID)

Het idee hierachter is dat niet-identificerende gegevens niet zelf hoeven te worden gemaskeerd, omdat ze hun gevoeligheid/vertrouwelijkheid verliezen als ze niet meer (vanwege het maskeren van de identificerende gegevens) kunnen worden herleid naar een persoon.

N.B. Niet alle identificerende velden kunnen worden gemaskeerd. Een geboortedatum, bijvoorbeeld, is wel identificerend, maar dusdanig belangrijk in de rekenregels van UWV dat maskering ervan de gemaskeerde waarde onbruikbaar zou maken voor gebruik in rapportage/analyse.
Dergelijke velden worden wél gedefinieerd als identificerend, maar de erop toegepaste

---

De GegevensLeveringsAfspraak (GLA) is het contract waarin de afspraken rondom een levering van gegevens aan een afnemer **uit** het DIM zijn vastgelegd.

Het GIA-proces wordt, ten tijde van het schrijven van dit ontwerp, aangepast door het Implementatieteam GD/DWH (geen onderdeel van het project DataFabriek).
Ook wordt daarbij gekeken in welke mate de Gegevensdiensten-applicatie UGC een rol kan spelen bij beheer/opslag van deze contracten.

GIA's en GLA's bevatten contract-metadata en compliance-metadata

RLO~~RLO~~)

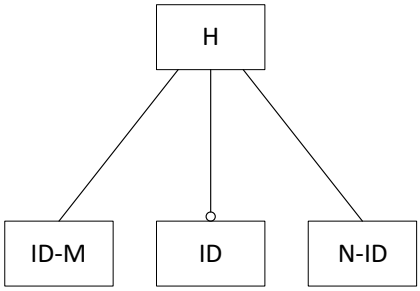
[50] Op dit type velden wordt vaak "row level masking" toegepast. Zie paragraaf 8.3.5~~8.3.6~~ (Maskeren).

maskering is "geen maskering". Het veld staat dus identiek in beide satellieten met identificerende data (ID en ID-M).

Elke tabel uit een bronlevering (zoals beschikbaar in de staging-laag) krijgt in de bronzone een eigen hub. Als meerdere tabellen details leveren voor hetzelfde bovenliggende object worden ze dus niet, zoals in standaard-DataVault2, verwerkt als meerdere satellieten van dezelfde hub.

Een tabel uit een bronlevering (zoals beschikbaar in de staging-laag) wordt in de bronzone altijd omgevormd tot een hub met drie bijbehorende satellieten. Op basis van of een attribuut is aangemerkt als een identificerend gegeven wordt bepaald hoe en in welke satelliet(en) deze zal worden opgeslagen. Identificerende gegevens worden altijd in twee van de satellieten opgeslagen namelijk in de gemaskeerde variant en in de ongemaskerde variant. Alle niet-identificerende gegevens zullen worden opgeslagen in de satelliet met de overige gegevens.[51] [52]

Alle vier deze tabellen worden op gelijke wijze geversioneerd (de hub ook!). De relatie tussen de Hub en de satellieten is dus niet, zoals inDataVault2, een 1-op-veel relatie, maar een 1:1 relatie. Hieronder is een schematische weergave van de relaties tussen de Hub en zijn bijbehorende satellieten.



De Hub en de satellieten vormen een eenheid; elke rij in de Hub heeft altijd precies één corresponderend record in elk van de drie (of twee[53]) bijbehorende satellieten. Dat betekent dat een wijziging van een brongegeven altijd betekent dat er een nieuwe versie komt van zowel de hub als van de satellieten, dus ongeacht of die hub of satelliet door de wijziging zelf geraakt wordt.

Hierbij wordt gebruik gemaakt van een begin- en een einddatum/tijd, en de oude versie wordt altijd afgesloten door middel van het invullen van een einddatum/tijd.

De versionering van de Hub vindt plaats omdat er in de hub ook, over tijd wijzigende, technische informatie staat die gebruikt wordt bij het laden van de bronzone.[54]

Door deze opzet:

- Is het eenvoudig om de updates te verwerken zonder ook nog te moeten bepalen in welke satelliet de update moet worden uitgevoerd.

[51] Deze opsplitsing van één satelliet naar drie satellieten wijkt af van de DataVault2 standaard. Het is toegepast om op een relatief eenvoudige wijze gemaskeerde en ongemaskerde gegevens eenduidig te kunnen verwerken, maar gescheiden te kunnen benaderen.

[52] Ook als de tabel uit de bronlevering géén identificerende attributen bevat worden er drie satellieten gebruikt. Twee daarvan (ID en ID-M) bevatten dan slechts technische velden. Deze benadering is gekozen om het aanmaken van ontkoppelviews volledig te kunnen standaardiseren.

[53] Na vernietiging van de identificerende gegevens heeft een hub-rij geen corresponderende rij meer in de ID-satelliet. Zie paragraaf 5.2~~5.2.2~~ (Twee typen Data Lifecycle Management) en 14.3~~15.3~~ (DLM in de bronzone~~DLM in de bronzone~~) voor meer informatie.

[54] Zie paragraaf 8.3.2(Bepaling delta's)

- Is het eenvoudiger om een performante ontkoppelview te maken. Alle hubs en satellieten hebben dezelfde administratieve tijdlijn (ook als er in slechts één satelliet inhoudelijk iets gewijzigd is). Ze zijn daardoor een-op-een te matchen zonder complexe matches te moeten uitvoeren.

8.2.2.2 Geen link-tabellen

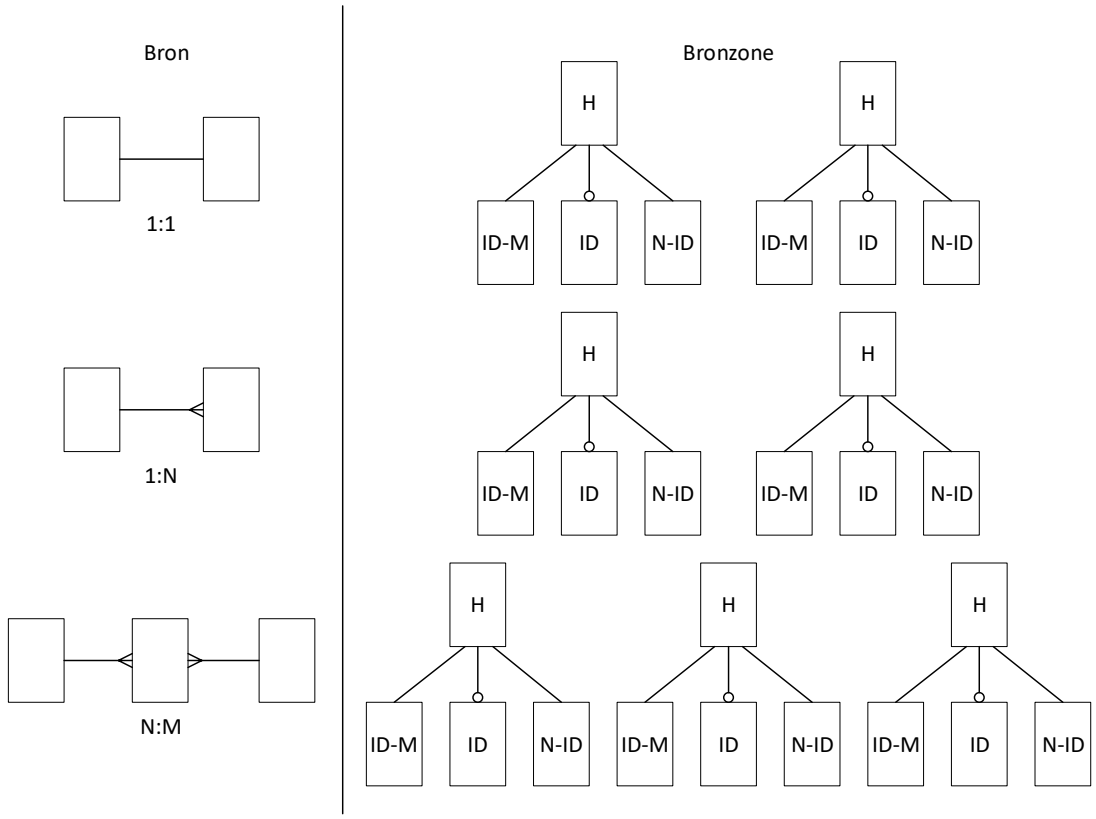
In de bronzone vindt **geen** integratie plaats.

Hierdoor wordt het mogelijk om alleen hubs en satellieten te gebruiken, en geen DataVault "links":

- Verwijssleutels in de brontabellen worden behandeld als "gewone", al dan niet identificerende, attributen, en worden niet opgeslagen in de vorm van een "link"-tabel.

- Ook tabellen die in de bron zelf een link-functie vervullen (als koppeltabel in een N op M relatie) worden in de bronzone behandeld als "gewone" tabellen.

Ter verduidelijking:
Indien er twee tabellen uit een bron worden geleverd met een 1 op 1 relatie zal elke tabel zijn eigen Hub met satellieten krijgen in de bronzone. Dit geld ook voor 1 op N relaties. Bij een N op M relatie zal behalve de 2 tabellen ook de koppeltabel aangeleverd worden. Deze drie tabellen krijgen elk hun eigen Hub met satellieten. Er zal hier geen gebruik gemaakt worden van link tabellen zoals gebruikelijk is in DataVault2.



8.2.2.3 Bedrijfssleutels vs. primaire sleutels

Een **bedrijfssleutel** is een unieke sleutel tot een gegeven in een tabel bestaande uit één of meer velden, die allemaal een (herkenbaar) bedrijfsgegeven zijn.

Een **technische sleutel** is unieke sleutel tot een gegeven in een tabel, meestal bestaande uit één technisch (gegenereerd, betekenisloos) veld, soms aangevuld met één of meer extra velden (meestal versiedatums). De technische sleutel wordt meestal gevuld met behulp van een database sequence.

Een **primaire sleutel** kan een bedrijfssleutel of een technische sleutel zijn, in beide gevallen is het een unieke sleutel tot een gegeven in een tabel.

In DataVault2 wordt een gehashte **bedrijfssleutel** afkomstig uit de bron gebruikt als primaire sleutel van de Hub. Het DIM gebruikt echter altijd de gehashte **primaire sleutel** afkomstig uit de bron (gecombineerd met de startdatum van de versie; de hub is immers geversioneerd) als primaire sleutel van de Hub, ook als die primaire sleutel een **technische sleutel** is.

Hierdoor blijft het in het DIM mogelijk om altijd de correcte historie van de administratieve werkelijkheid van een bron vast te leggen, ook in het geval dat, bijvoorbeeld, de bedrijfssleutel in de bron aanpasbaar is. Binnen het DIM noemen we deze primaire sleutel van de hub de **hub-sleutel**.

Een nadeel van een gehashte hub-sleutel is dat deze vrij lang kan worden. Het koppelen op dergelijke erg grote velden is in databases erg inefficiënt omdat de koppelvelden gematched moeten worden; twee numerieke velden zijn sneller te vergelijken dan twee lange strings.
Bij het bij elkaar brengen van de hub en de satellieten in een query kan dit dus performance-problemen opleveren. Dit is met name in het geval voor hub/satelliet-structuren met veel gegevens.

Het voorgaande nadeel wordt in het DIM opgelost door de introductie van een unieke DIM-sleutel. Deze staat in de Hub en de Satellieten. De DIM-sleutel is een op basis van een sequence gegenereerde sleutel die ervoor zorgt dat de administratieve versie van een hub en zijn bijbehorende satellieten eenvoudig en performant aan elkaar te koppelen zijn.

Zie hoofdstuk 15 (Generieke oplossingen~~Generieke oplossingen~~) voor meer informatie over de algemene opzet van harnassen. Gedetailleerde beschrijvingen zijn opgenomen in de documentatie van de agile bouw teams. Op dit moment is de realisatie bezig[GL(6)]*, de betreffende documentatie is aan veranderingen onderhevig.*

8.2.2.4 Versionering o.b.v. administratieve tijdlijn

De versionering in de bronzone-data is gebaseerd op de administratieve tijdlijn[55] van de bron[56]. Alle overige door de bron aangeleverde historie-informatie wordt door het DIM als functionele attributen gezien. De administratieve tijdlijn van het DIM zelf (de verwerkings-momenten) wordt in de vorm van technische velden aan de (geversioneerde) brongegevens toegevoegd.

Bij deze versionering wordt zoveel mogelijk gebruik gemaakt van in de bron beschikbare wijzigingsinformatie. Zie paragraaf 8.3.1 (Versionering in de bronzone-data~~Versionering in de bronzone-data~~) en Bijlage D: Volgen administratieve historie bron *voor meer informatie.*

8.2.2.5 Eigenaarschap

Van alle gegevenselementen in de **bronzone** dient duidelijk te zijn wat ze betekenen:

[55] Binnen UWV ook wel de "transactie-dimensie" genoemd. Zie https://digitalewerkplek.sharepoint.uwv.nl/documentcenter/Documenten/Gegevensdiensten/Beleidsdocument/Beleid%20Tijdsdimensies%20versie%201.4.pdf voor meer informatie.

[56] Of een benadering daarvan. Zie paragraaf 8.3.1 (Versionering in de bronzone-data)

- Ieder gegevenselement in de **bronzone** moet (o.b.v. de informatie in de RLO) gekoppeld zijn aan een functionele term afkomstig uit de functionele terminologie-laag in IGC.
- In deze koppeling moet ook de eventuele impact van de historische verwerking in het DIM zijn opgenomen, bijvoorbeeld of de administratieve tijdlijn is overgenomen uit de bron of afgeleid binnen het DIM.

8.3 Laadlogica bronzone

De gegevens in de **staging-laag** zijn, qua structuur, onafhankelijk van de wijze van levering; alle leverings-specifieke zaken zijn namelijk al uitgevoerd in de **ontsluitingszone**.[57]

Daarnaast bouwt de **laadlogica bronzone** een hub met satellieten in stappen op.

Het verwerken van de gegevens in de **staging-laag** in bronzone-data kan daardoor op één manier, en door één harnas worden uitgevoerd: het bronzone-harnas.

8.3.1 Versionering in de bronzone-data

De versionering in de bronzone-data gebruikt de in de bron gebruikte primaire sleutel (deze kan dus technisch zijn) als basis, omdat deze sleutel de werkelijk in de bron gebruikte granulariteit van het object representeert. De versionering zelf gebeurt op basis van de administratieve tijdlijn, en sluit zo nauw mogelijk aan op die administratieve tijdlijn in de bron:

- Als de bron een administratieve tijdlijn heeft dan wordt deze gebruikt.
- Als de bron geen enkele administratieve tijdlijn heeft dan wordt het extractiemoment van de bronlevering omgevormd tot de administratieve tijdlijn.
- Als een bron alleen transacties heeft is de transactie-tijdlijn de administratieve tijdlijn. Indien transacties overschreven worden voordat deze met behulp van een bronlevering geleverd zijn aan het DIM zal er hier verlies van historie optreden doordat het DIM deze historie nooit aangeleverd heeft gekregen.

N.B. De actuele versie van een gegeven krijgt in het DIM altijd als einddatum "31-12-9999", ook als de bron hier een andere conventie voor gebruikt.

Alle overige door de bron aangeleverde historie-informatie wordt door het DIM als functionele attributen gezien.

Op basis van de RLO wordt bepaald hoe de administratieve tijdlijn van de betreffende hub en satellieten combinatie in het DIM gevuld wordt. Dit zal in de stuur-metadata worden vastgelegd.

Voor het opbouwen van de historische afspiegeling van de werkelijkheid van de bron zijn er verschillende scenario's met betrekking hoe data wordt aangeleverd. Hoe data wordt aangeleverd heeft impact op de stuur-metadata welke gebruikt wordt om de data te verwerken van de staging laag naar de bronzone.

Doordat we van de bronlevering en elke tabel in de bronlevering weten hoe we de administratieve tijdlijn kunnen bepalen is het mogelijk om een tussenstap van de staging in naar de bronzone te doen. Dit doen we door de **staging in** tabellen om te vormen naar de hub/satelliet-structuur zoals deze in de bronzone gebruikt wordt. Deze tabellen noemen we de **staging uit** tabellen. In deze stap zijn zaken als afgeleide velden en maskering al uitgevoerd.

Op dit punt is het mogelijk geworden om voor alle leveringswijzen van de historie (door de bron) en elke mate van beschikbaarheid van de administratieve historie (in de bron) door middel van een

[57] Meer informatie over de staging-laag in paragraaf 7.3 (Staging-laag)

delta-vergelijking te bepalen wat er met de data gedaan moet worden ten opzichte van de bestaande situatie in de bronzone.

Voor zowel een incrementele als een stapelbare bronlevering zal dit alleen leiden tot toevoegingen van nieuwe rijen in de bronzone en wijzigingen van bestaande rijen (vaak is dit alleen zetten van een einddatum/tijd van de voorgaande versie).
Voor een volledige bronlevering zal dit kunnen leiden tot toevoegingen van nieuwe rijen en wijzigingen van bestaande rijen. Maar doordat we in de bronzone ook kunnen controleren of er gegevens verwijderd zijn in de bron kan het ook leiden tot het verwijderen van rijen in de ongemaskeerde [GL(7)identificerende satelliet (ID).

Verdere detail-uitwerking en eventuele uitbreiding van bovenstaande punten worden gedaan in het technisch ontwerp van de ETL-bouwblokken en harnassen. Daar het project agile werkt zijn deze technische detailleringen en keuzes onderdeel van de deliverables van de betreffende agile teams.

8.3.2 Bepaling delta's

Om op record-niveau te kunnen bepalen wat de delta is tussen de hub/satelliet-combi's in **staging uit** (afkomstig uit de bronlevering) en de hub/combi's die al in de **bronzone-data** aanwezig zijn is er een beslisboom te maken.

Deze beslisboom bestaat uit twee delen.

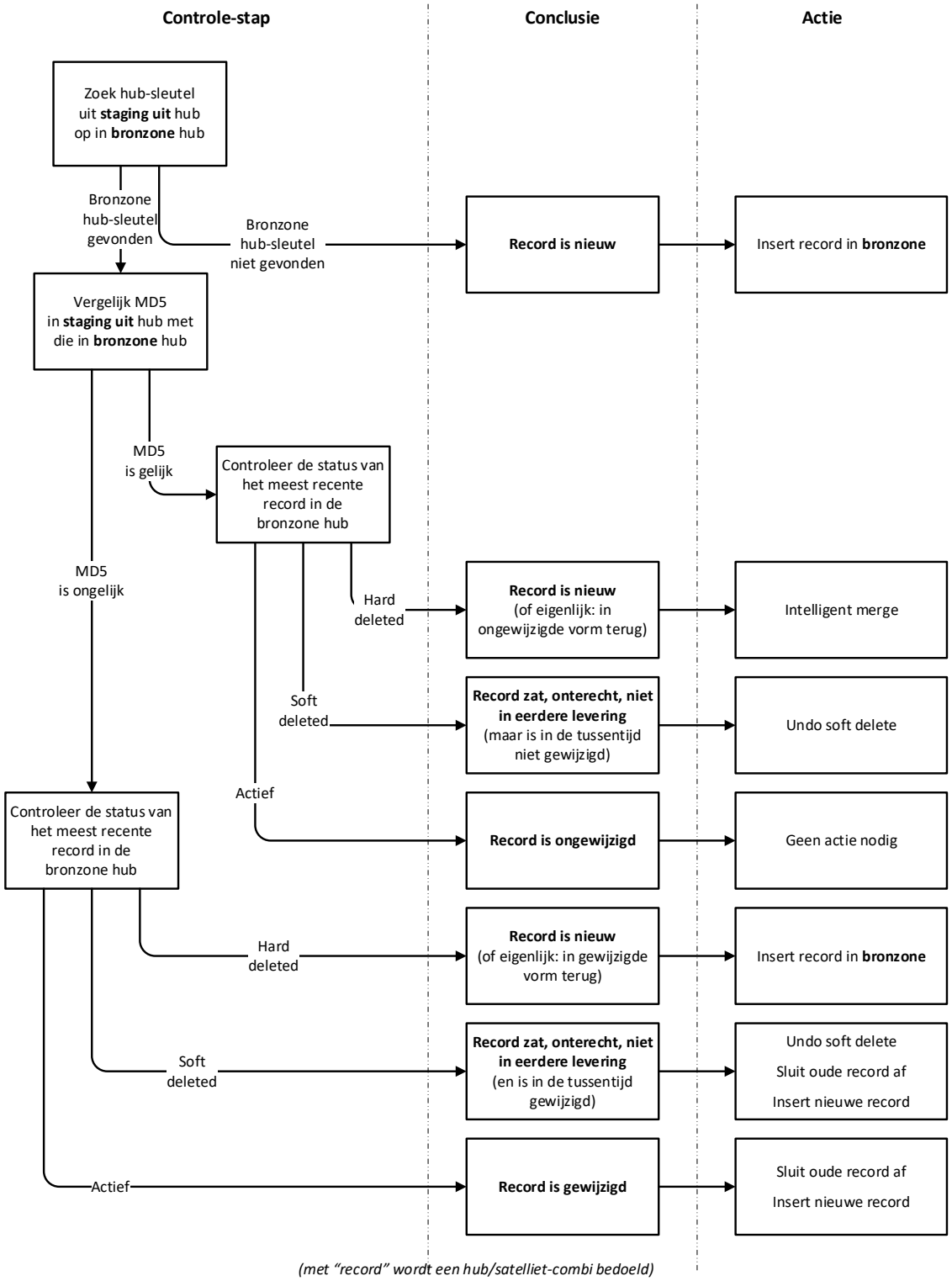
- Vanuit staging uit naar de **bronzone-data** kijken (altijd)
- Vanuit de **bronzone-data** naar **staging uit** kijken (alleen mogelijk als een tabel volledig wordt aangeleverd)

Voor alle tabellen geldt dat er vanuit de hub in **staging uit** gekeken wordt naar de hub in de **bronzone-data**. Door de md5-berekening die we in beide hubs beschikbaar hebben kan makkelijk gecontroleerd worden of er voor de bestaande hub-sleutel iets gewijzigd is.

Het delta-proces is zo ingericht dat:

- het ook om kan gaan met bronnen die zelf al geversioneerd zijn op basis van de administratieve tijd;
- het ook om kan gaan met bronnen die regulier een incrementele bronlevering doen, en in een lagere frequentie een volledige (t.b.v. reconciliatie en notificatie van verwijderingen)

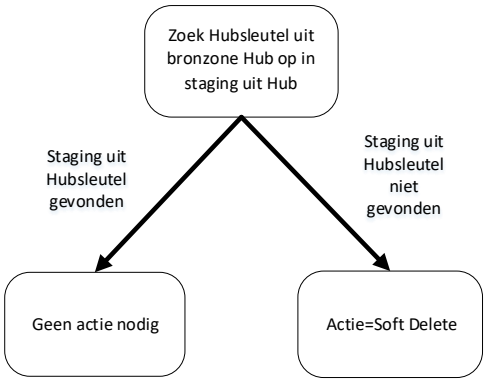
Let op! Voor tabellen met een (volledige of beperkte) administratieve tijdlijn moet het veld dat deze tijdlijn representeert (de startdatum/tijd van de administratieve versie respectievelijk de "datum/tijd laatst gewijzigd") bij het matching proces meegenomen worden met de hub-sleutel. Lees in dat geval in de beslisboom hieronder voor "Hubsleutel" steeds "Hubsleutel plus veld administratieve tijdlijn". In alle andere gevallen wordt er in de bronzone naar de meest recent opgebouwde administratieve versie (met deze hub-sleutel) gekeken.



(met "record" wordt een hub/satelliet-combi bedoeld)

Voor tabellen welke volledig worden aangeleverd is het mogelijk om vanuit de hub in de **bronzone-data** te checken welke records niet meer beschikbaar zijn in de hub van de **staging uit**.

Dit wordt gedaan door middel van de Hub sleutel.



Als resultaat van de delta-bepaling is voor alle rijen in de **staging uit** hub/satelliet-combi's bekend wat de bijbehorende acties zijn voor de verwerking naar de hub/satelliet-combi's in de **bronzone-data**. Deze verwerking kan dan ook in de vervolgstap uitgevoerd worden.

8.3.3 Volgen van het DLM van de bron: harde en zachte verwijderingen

De bron is leidend bij het Data Lifecycle Management (DLM) van de ongemaskeerde identificerende gegevens in de bronzone.[58]

Om bij onvoorziene problemen van verwerkte bronleveringen niet gelijk alle historie kwijt te zijn worden gegevens eerst zacht verwijderd. Dit betekent dat deze data in de ongemaskeerde ontkoppelview niet meer zichtbaar is, maar in de bronzone in de ongemaskeerde (ID) satelliet nog wel bestaat.

Het hard verwijderen wordt in een separaat schoningsproces gedaan, op basis van een zogenaamde "grace periode"; na afloop van die periode wordt de zachte verwijdering omgezet in een hard (fysieke) verwijdering.

Door eerst zacht te verwijderen is het mogelijk om gedurende de grace periode per ongeluk verwijderde gegevens zonder gegevensverlies weer te herstellen.
De grace periode wordt, per bron, in de stuur-metadata vastgelegd op basis van afspraken in de GLO. Indien een bron niets afspreekt zal deze grace periode 90 dagen zijn. Het is ook mogelijk om deze grace periode 0 te maken.

In meer detail werkt dit proces als volgt:

- Als, o.b.v. van een bronlevering, wordt gedetecteerd dat een gegeven fysiek is verwijderd uit de bron, dan krijgen de actuele versies van dat gegeven in de bronzone (in hub en satellieten) het verwijdermoment als einddatum (i.p.v. de defaultwaarde 31-12-9999) én wordt de **DIM status** van het gegeven (in alle versies van de hub) omgezet naar "zacht verwijderd".
- In de ongemaskeerde ontkoppelviews wordt vervolgens geen enkele versie meer getoond. Er is vanuit gebruikersperspectief dus geen verschil tussen een zachte en een harde verwijdering.
  De andere ontkoppelviews worden niet beïnvloed; alleen de ongemaskeerde ontkoppelviews worden niet gefilterd op de DIM status.
- Mocht binnen de grace periode de betreffende data toch weer aangeleverd worden dan wordt de DIM status van alle zacht verwijderde hub-versies weer leeg gemaakt.

[58] Voor de gemaskeerde en overige gegevens geldt een ander DLM. Zie paragraaf 14.3 (DLM in de bronzone~~DLM in de bronzone~~) voor meer details.

- Als de grace periode verstreken, dan worden voor alle als "zacht verwijderd" gemarkeerde hub-versies de bijbehorende ongemaskeerde satelliet-versies verwijderd, en wordt volgens de DIM status in de hub-versies op "hard verwijderd" gezet.

Let op! Het DIM is wel afhankelijk van het feit dat we weten dat een gegeven uit een bron is verwijderd. Hiervoor zijn de volgende mechanismen voorzien:

- Volledige gegevensleveringen.
  Hieruit kan door vergelijking met de inhoud van de bronzone bepaald worden welke gegevens in de bron verwijderd zijn.
- Expliciete aanlevering van het feit dat een gegeven verwijderd is in de bron.
- Binnen het DIM naspelen van de verwijdering.
  De DLM-logica van de bron wordt, o.b.v. door de bron geleverde specificaties, binnen het DIM nagebouwd en uitgevoerd.[59]

N.B. Het gebruik van zachte verwijderingen met een grace periode is een "vangnet" voor leveringsfouten die eigenlijk niet mogen voorkomen. Ongedaan maken van deze zachte verwijderingen kan verderop in het DIM (in integratiezone en bedrijfszone) aanzienlijke doorlooptijd vergen. Het is dus belangrijk om onvolledige bronleveringen zoveel mogelijk al "af te vangen" door pakbon-controles en dergelijke, zodat ze niet verder verwerkt worden.

8.3.4 Aanmaken van afgeleide velden

De bronzone volgt de administratieve werkelijkheid van de bron; de transformaties binnen de bronzone zijn dus beperkt tot het omvormen van de brongegevens naar een hub/satelliet-structuur en het maskeren van identificerende velden.

In een aantal gevallen is het noodzakelijk om hier van af te wijken door toch al in de bronzone afgeleide velden op te slaan.

Het maken van afgeleide velden wordt in de stuur-metadata ingericht door aan te geven:

- dat het veld een afgeleid veld is;
- wat er gedaan moet worden om de ongemaskeerde versie van het afgeleide veld tot stand te brengen
- in welke vorm dit afgeleide veld daarna (eventueel) gemaskeerd dient te worden.

Afgeleide velden mogen in de bronzone alleen gebruikt worden als:

a) De afleiding noodzakelijk is voor een proces binnen de bronzone (bijvoorbeeld maskeren), of
b) De afleiding bron-gerelateerd is (en ook door de bron gespecificeerd en gevalideerd), en noodzakelijk is voor (vrijwel) alle verdere gebruik van de gegevens.

Alle andere afleidingen dienen in de **integratiezone** of **bedrijfszone** plaats te vinden.

Ad a)
Op dit moment zijn hiervan twee gevallen onderkend:

- Velden die moeten worden geüniformeerd voordat ze gemaskeerd kunnen worden.
- Velden die op twee verschillende manieren gemaskeerd moeten worden

Zie paragraaf 8.3.5 (Maskeren) voor een uitwerking van beide gevallen.

Ad b)

[59] In de GLO worden de verantwoordelijkheden voor dat naspelen (de bron specificeert, valideert, accepteert) expliciet vastgelegd.

Op dit moment zijn hiervan nog geen gevallen onderkend.

8.3.5 Maskeren

Zie hoofdstuk 5 (Maskering en DLM) voor algemene informatie over maskeren.

De op een attribuut uit te voeren maskering is in de stuur-metadata vastgelegd als **maskeringsklasse**. Voor elke van deze maskeringsklassen wordt een specifiek ETL-bouwblok (de maskeringsbouwsteen) ontwikkeld.[60]

Als van een veld is aangegeven dat het identificerend is dan komt het ongemaskeerd in de ongemaskeerde (ID) satelliet, en gemaskeerd in de gemaskeerde (ID-M) satelliet.

Voor alle velden welke identificerend zijn zal er een maskering worden opgegeven. Voor identificerende velden waarvoor de vereiste maskering "geen maskering" blijkt te zijn is de waarde in de gemaskeerde (ID-M) en de ongemaskeerde (ID) satelliet dus gelijk.[61]

Sommige vormen van maskeren vereisen meerdere stappen:

- rij-niveau maskering (row level masking)
- maskering na uniformering
- meerdere maskeringen op hetzelfde veld

De aanpak van deze vormen van maskeren wordt in onderstaande paragrafen verder uitgewerkt.

8.3.5.1 Rij-niveau maskering (row level masking)

Voor sommige brongegevens kan de betekenis ervan per rij verschillen, bijvoorbeeld als het, in de bron, als een "generiek" veld is gedefinieerd, en de betekenis ervan (en dus ook of het identificerend is of niet) vastgelegd is in een ander veld. Dergelijke velden worden in de bronzone beschouwd en verwerkt als "identificerend", maar de uit te voeren maskering kan per rij verschillen.

Een voorbeeld hiervan is de bron PEERCODE, die ingevulde vragenlijsten aan het DIM levert, waarbij één veld het antwoord bevat, en één veld een code voor de vraag waar het antwoord bij hoort. Ook levert PEERCODE metadata op die vraag-codes aan betekenis (adres, telefoonnummer, etc) en maskeringsklasse koppelt.

In de bronzone zou een dergelijke maskering als volgt kunnen worden uitgevoerd:

- Het veld krijgt in de stuur-metadata een maskeringsklasse waaruit blijkt dat je row level masking moet uitvoeren, bijvoorbeeld: "Rij-maskering ZW-antwoorden".
- De maskeringsbouwsteen die bij "Rij-niveau maskering ZW-antwoorden" hoort voert het rij-niveau maskeren uit, in dit geval door o.b.v. door PEERCODE geleverde info een

[60] Zie paragraaf 15.3 (Maskeringsbouwstenen) voor meer info.

1.1.2 [61] Zie ook paragraaf 8.2.2~~8.2.3~~ (Datamodel

Het datamodel van de bronzone is "op maat ontworpen" om gemaskeerde en ongemaskeerde gegevens met minimale gegevensduplicatie toch maximaal te scheiden. In het model worden (varianten op) een aantal uit DataVault2 ontleende concepten en begrippen toegepast, zoals hub/satelliet-structuren en hashkeys.
De bronzone is echter uitdrukkelijk géén DataVault2-implementatie; het datamodel is primair gebaseerd op de behoeften van het DIM, en niet op de Datavault2-theorie.
De volgende paragrafen beschrijven het datamodel, inclusief overeenkomsten en verschillen met DataVault2, in meer detail.
Hubs en satellietenHubs en satellieten)

maskeringsklasse op rij-niveau te bepalen (kan ook "geen maskering" zijn). Het ideaal is als dit "gewone" maskeringsklassen zijn (dezelfde die je gebruikt bij gewoon maskeren, op attribuut-niveau, dus "BSN", "telefoonnummer", etc), en dat ook de, nu per rij aangeroepen, maskeringsbouwstenen de "gewone" (attribuut-niveau) maskeringsbouwstenen zijn, of tenminste dat, voor dezelfde maskeringsklasse, attribuut-niveau maskeren gegarandeerd dezelfde maskeringslogica uitvoert als rij-niveau maskeren.

Bovenstaande als voorbeeld; het ontwerpproces loopt nog[GL(8]*, en kan resulteren in een net iets andere aanpak..*

N.B. Dit betekent dat er in de (ID-M) satelliet dus attributen zijn die in sommige rijen wel, en in andere rijen niet gemaskeerd zijn.

8.3.5.2 Maskering na uniformering

Een voorbeeld hiervoor is het BSN nummer dat binnen UWV soms met en soms zonder een voorloopnul wordt gebruikt. Dit leidt bij maskeren tot verschillende resultaten waardoor er geen integratie mogelijk is voor het gegeven. Door in het afgeleide veld een uniforme versie van het BSN te maken en deze daarna te maskeren is zowel de ongemaskeerde als de gemaskeerde variant koppelbaar.
Om de traceerbaarheid van de gegevens ondanks deze uniformering in stand te houden wordt hier gebruik gemaakt van afgeleide velden[62].

De eerste stap is dan het veld uniform te maken waarna het alsnog gemaskeerd wordt. De (ID-M) en (ID) satellieten bevatten dan zowel het origineel als het afgeleide veld. Het originele veld is bijvoorbeeld slechts in beperkte mate koppelbaar waar het afgeleide veld wel koppelbaar is.

De doorlopen stappen zijn:

- Maak een afgeleid veld met daarin de geüniformeerde waarde
- Sla in de ongemaskeerde satelliet zowel het oorspronkelijke als het afgeleide (geüniformeerde) veld op
- Sla in de gemaskeerde satelliet het oorspronkelijke veld "verborgen" (uitgekruist of leeggelaten) op, en het afgeleide veld "koppelbaar vervangen".

8.3.5.3 Meerdere maskeringen op hetzelfde veld

Een voorbeeld hiervan is de postcode. In het overgrote deel van de analyses/rapportages voldoet het postcodegebied (de eerste vier cijfers), maar soms is rapportage per UWV-regio noodzakelijk. Helaas zijn er postcodegebieden die deels in de ene, en deels in een andere UWV-regio vallen.

Om te voorkomen dat de postcode daarom helemaal niet gemaskeerd kan worden zijn er tenminste twee alternatieven[63]:

- Zowel postcodegebied als een "koppelbaar vervangen" gemaskeerde versie van de postcode zelf ter beschikking stellen in de bronzone.[64]
- Zowel postcodegebied als UWV-regio beschikbaar stellen in de bronzone.

[62] zie paragraaf 8.3.4 (Aanmaken van afgeleide velden)
[63] Er zijn voor dit voorbeeld meerdere oplossingen mogelijk, allemaal gebaseerd op het slim inzetten van afgeleide velden. De hier beschreven alternatieven zijn slechts bedoeld om het concept te verduidelijken; de definitieve oplossing wordt bij het detailontwerp van de maskeringsbouwstenen bepaald.
[64] De UWV-regio kan dan, door ook een gemaskeerde koppeltabel postcode vs. UWV-regio in de bronzone op te nemen (geleverd door de UWV-bron voor deze referentiegegevens), in de integratiezone worden afgeleid.

N.B. Voor beide alternatieven is een "koppeltabel" tussen postcode en UWV-regio noodzakelijk. Deze dient door de UWV-bron voor deze referentiegegevens aan het DIM geleverd te worden. Binnen het DIM zelf worden immers geen gegevens beheerd.

Voor het eerste alternatief zijn de te doorlopen stappen:

- Maak een afgeleid veld voor postcodegebied.
- Sla in de ongemaskeerde satelliet zowel het oorspronkelijke veld (de postcode) als het afgeleide veld (het postcodegebied) op
- Sla in de gemaskeerde satelliet het oorspronkelijke veld (de postcode) "koppelbaar vervangen" op, en het afgeleide veld (het postcodegebied) "niet gemaskeerd".
- Stel ook de "koppeltabel" met de UWV-regio ter beschikking in de bronzone, met in de gemaskeerde versie de postcode "koppelbaar vervangen" gemaskeerd.
  De UWV-regio kan dan ook voor gemaskeerde gegevens in de integratiezone worden afgeleid.

Voor het tweede alternatief zijn de te doorlopen stappen:

- Maak een afgeleid veld voor UWV-regio.
- Sla in de ongemaskeerde satelliet zowel het oorspronkelijke veld (de postcode) als het afgeleide veld (de UWV-regio) op[65]
- Sla in de gemaskeerde satelliet het oorspronkelijke veld "verborgen" (letters uitgekruist) op, en het afgeleide veld "niet gemaskeerd".

Het tweede alternatief vereist bedrijfslogica in de bronzone. Het eerste alternatief heeft dus de voorkeur. [GL(9]Bijkomend voordeel van het eerste alternatief is dat deze ook om kan gaan met wijzigende regio-indelingen.

8.4 Bronzone-harnas

Er is ervoor gekozen om op alle gegevens in de **staging-laag** een delta-bepaling uit te voeren, ook als dat, bijvoorbeeld bij transactiegegevens, strikt genomen niet noodzakelijk is. Dit maakt het mogelijk één harnas te gebruiken. De proces-flow van dit harnas is dan relatief simpel. De meeste complexiteit gaat zitten in de bepaling van wat moet er gedaan worden met een ontvangen gegeven.

Bij de start van het harnas wordt er vanuit gegaan dat alle data-issues welke verwerking in de weg kunnen staan al in het voorgaande harnas (in de ontsluitingszone) gedetecteerd zijn. In het bronzone-harnas kunnen er dus enkel technische fouten optreden.

Vanuit het oogpunt van technische fouten (bijvoorbeeld schema vol) is het harnas herstartbaar omdat de al verwerkte data door de delta-bepaling als ongewijzigd wordt gezien, en dus niet opnieuw verwerkt wordt.

Voor elke bronlevering doorloopt het bronzone-harnas dezelfde stappen:

1 Start

De **master sequence bronverwerking**[66] checkt, op basis van de returnstatus van de laadlogica staging, of de verwerking tot dit punt correct is verlopen, start vervolgens het bronzone-harnas, en geeft daaraan ook zijn inputparameters (Run ID, Bronlevering ID, Bronlevering Naam, Bron Naam, Procesdatum) mee.

[65] De afleiding van de regio vereist dezelfde tabel met referentiegegevens als die in het eerste alternatief.

[66] Zie paragraaf 8.6 (Master Sequence bronverwerking) voor meer info.

Verder wordt er ook een omgevingsafhankelijke default parameter-set meegeven. De inhoud van deze parameter-sets is gebaseerd op de omgevingsvariabelen die op DIM project-niveau in DataStage gedefinieerd zijn (zie paragraaf 8.6.1).

Haal, m.b.v. de door de master sequence meegegeven parameters, de voor de verwerking van de bronlevering benodigde stuur-metadata op.

Maak een record aan in de log-metadata met als status "lopend".

2 Initialisatie
Bepaal, op basis van de DataStage omgevingsvariabelen, de omgevingsafhankelijke[67] delen van paden en eventuele bestandsfolders.

3 Voor alle tabellen van de bronlevering in **staging in**:

3.0 (preprocessor)
Als de standaard-generatie (in 3.1) niet voldoet, zet in deze stap dan de afwijkende SQL-statements klaar. Deze worden dan gebruikt in plaats van de standaard SQL-statements uit stap 3.1

3.1 Genereer (m.b.v. DataStage[68]), op basis van de stuur-metadata, de DDL- en SQL-statements om de hub/satelliet-combi's in **staging uit** te creëren en te vullen (inclusief technische velden, afgeleide velden en maskering).

3.2 Vul de hub/satelliet-combi's in **staging uit** door deze DDL- en SQL-statements uit te voeren

3.3 Bepaal op basis van de delta beslisboom (zie paragraaf 8.3.2) welke actie er voor elke rij in de hub/satelliet-combinatie uitgevoerd moet worden

3.4 Doe, op basis van de net bepaalde acties, een "intelligente merge" / update in de hub/satelliet-combi's in de **bronzone-data**, vanuit de hub/satelliet-combi's in **staging uit**.

X1 Verzend notificaties
Meld succes of issues via e-mail aan de relevante partijen.
Welke meldingen naar welke partijen moeten, en welke emailadressen voor deze partijen gebruikt moeten worden, staat in de stuur-metadata.

Op dit moment krijgt alleen DIM-beheer notificaties. In fase 5 worden de notificatie-vereisten verder uitgewerkt, en wordt deze stap (en de bijbehorende stuur-metadata) aangepast om die vereisten te ondersteunen.

X2 Einde job
Vervang de log-markering "lopend" door "klaar" of "afgebroken", afhankelijk van het resultaat van de verwerking.
[einde ETL-job]

[67] Met "omgevingsafhankelijk" wordt bedoeld: afhankelijk van de OTAP-omgeving waarin het proces draait.
[68] DataStage genereert die statements op basis van het in de DataStage-metadata vastgelegde doel-RDBMS (op dit moment dus Oracle).

N.B. Elke (sub)stap slaat de uitkomsten van de(sub)stap, en eventuele overige relevante procesinformatie, op in de **log-metadata**. Als de (sub)stap een fout detecteert, dan wordt, na logging, de verdere verwerking afgebroken en direct doorgesprongen naar de afronding van het proces (stap X1).

Bovenstaande geeft een globale indruk van het bronzone-harnas. Voor meer details wordt verwezen naar de documentatie welke opgesteld is/wordt tijdens de realisatie.
N.B. Op dit moment is de realisatie bezig, de betreffende documentatie is dus nog aan veranderingen onderhevig.

8.5 Ontkoppelviews bronzone

De ontkoppelviews in de bronzone dekken altijd de gegevens van precies één bron-entiteit, en dus ook van precies één hub/satelliet-structuur in de bronzone-data.
In de views wordt deze hub/satelliet-structuur overigens verborgen door alle delen van die structuur aan elkaar te joinen en zo "plat te slaan".

Zoals in paragraaf 8.1 vermeld dienen de ontkoppelviews vier doelen:

1. Ze maken de technische structuur van de bronzone-data onzichtbaar voor de achterliggende zones; die zien géén hub/satelliet-structuren, maar "herkenbare", al dan niet geversioneerde, brondata.
2. Ze splitsen gemaskeerde en ongemaskeerde data; de achterliggende zones (en de afnemers daarvan) zien óf gemaskeerde óf ongemaskeerde persoonsgegevens (en in sommige gevallen geen van beide).
3. Ze verbergen de "zachte verwijderingen" voor het achterland; in de ontkoppelviews worden "zacht verwijderde" gegevens namelijk niet getoond. Doordat de laadlogica voor de **integratiezone en/of bedrijfszone** zich baseert op de ontkoppelviews zullen "zachte verwijderingen" dus als "harde verwijderingen" doorwerken naar de **integratiezone** en de **bedrijfszone.**
4. Ze ontkoppelen het verversen van de bronzone-data van het gebruiken ervan; achterliggende zones (en de afnemers daarvan) blijven tijdens verversen de "oude" situatie (voor verversen) zien. Na verversen "flipt"[69] dit in één keer over naar de nieuwe, volledig ververste, situatie.

Daarnaast maken ontkoppelviews het mogelijk om, met behulp van parameters, gemaskeerde en ongemaskeerde varianten van een informatiegebied (in de integratiezone) of informatieproduct (in de bedrijfszone) te laden met hetzelfde stuk maatwerk-ETL.

Met name vanwege punt 2 loopt toegang tot de bronzone-data **altijd** via de ontkoppelviews (ook voor de laadprocessen van integratiezone en bedrijfszone), en toegang buiten de ontkoppelviews om zal als een (potentieel) datalek moeten worden beschouwd.
Enige uitzondering daarop is toegang door DIM-beheer (t.b.v. fout zoeken) in geval van incidenten.

Daarnaast mag een ontkoppelview nooit de mogelijkheid bieden om een link tussen gemaskeerde en ongemaskeerde data te leggen (of zelf die twee typen data combineren).
Dit om te voorkomen dat de maskering eenvoudig ongedaan kan worden gemaakt, waardoor er een, binnen de AVG, onacceptabel privacy risico kan ontstaan.

[69] Zie paragraaf 8.5.4 (Materialiseren ontkoppelviews en "flip")

8.5.1 Zes varianten ontkoppelviews

De ontkoppelviews bestaan in zes varianten[GL(10)], verdeeld over drie paren:

	Gemaskeerde persoonsgegevens	Ongemaskeerde persoonsgegevens	Overige gegevens
Historie	**volledige gemaskeerde administratieve historie** van één bronobject met persoonsgegevens	**volledige ongemaskeerde administratieve historie** van één bronobject met persoonsgegevens	**volledige administratieve historie** van één bronobject (dat niet als "persoonsgegeven" hoeft te worden beschouwd
Actueel	**actuele gemaskeerde administratieve situatie** van één bronobject met persoonsgegevens	**actuele ongemaskeerde administratieve situatie** van één bronobject met persoonsgegevens	**actuele administratieve situatie** van één bronobject (dat niet als "persoonsgegeven" hoeft te worden beschouwd

De actuele views zijn eigenlijk "views op views"; ze filteren uit de bijbehorende (gemaskeerde, ongemaskeerde, of niet aan personen gerelateerde) view mét historie alleen de actuele waarden.

De term "actueel" moet hierbij gezien worden vanuit het perspectief van de administratieve historie. Als een bron, bijvoorbeeld, een geschiedenis van woonadressen van een werkzoekende bijhoudt, dan tonen de actuele ontkoppelviews niet het laatste adres, maar de meeste recent bijgewerkte gegevens per adres.

N.B. Als de bron géén administratieve historie bijhoudt, en het adres van de werkzoekende dus, in de bron, bij elke verhuizing wordt overschreven, dan zijn, in bovenstaande voorbeeld, "laatste adres" en "meeste recent bijgewerkte gegevens per adres" gegevens-technisch equivalent. De actuele ontkoppelviews tonen dan dus de facto wél het laatste adres.

Voor bron-entiteiten die identificerende attributen bevatten worden de eerste twee view-paren aangemaakt (maar niet die met "overige gegevens"), voor bron-entiteiten die géén identificerende attributen bevatten wordt juist alleen het derde view-paar ("overige gegevens") aangemaakt.

De gemaskeerde en de ongemaskeerde ontkoppelviews zijn, per paar, identiek in naam en structuur. Alleen het schema waarin ze staan verschilt.[70]
Van de ontkoppelviews met overige gegevens is er maar één paar; die views bevatten geen identificerende attributen, en dus is er geen sprake van (on)gemaskeerde kolommen.

8.5.2 Geen logica in ontkoppelviews

De ontkoppelviews hebben als enig doel ontkoppelen; alle transformaties/integraties anders dan die van "hub/satelliet" naar "platgeslagen" zijn óf al eerder in de bronzone uitgevoerd, óf worden dat pas in een achterliggende zone (integratiezone of bedrijfszone).

Opnemen van logica in de ontkoppelviews resulteert namelijk in**:**

- Transformatielogica op twee plekken en met twee technologieën, en daardoor complexer beheer, zowel qua kennis als omdat wijzigingen potentieel logica op meer dan één plek raken.
- Potentieel een breuk in de horizontale lineage, of meer inspanning om die breuk te voorkomen

[70] Zie paragraaf 8.8.1 (Database-inrichting) en 8.5.3 (Eén naam, twee schema's)

- Potentieel een lagere performance bij opbouwen van de view, met name bij grote tabellen.

In een ontkoppelview zal de hub met twee van de drie satellieten gecombineerd worden. De satelliet met de niet-identificerende gegevens is hiervan altijd een onderdeel. Voor de gemaskeerde ontkoppelviews zal de gemaskeerde satelliet de tweede satelliet zijn. Voor de ongemaskeerde ontkoppelviews zal de ongemaskeerde satelliet de tweede satelliet zijn.
De koppeling tussen de gemaskeerde/ongemaskeerde satelliet en de andere satelliet is daarbij een inner join. Hierdoor heeft verwijdering van gegevens in de bron (zodra dit verwerkt is in het DIM) ook tot gevolg dat de gegevens uit de ongemaskeerde ontkoppelview verdwenen zijn.

N.B. Voor "overige gegevens" bevatten zowel de gemaskeerde als de ongemaskeerde satelliet slechts technische velden. De ontkoppelviews op deze gegevens zijn, óf slechts gebaseerd op de hub plus één satelliet, of toch, vanwege verwerkingsstandaardisatie, op dezelfde satellieten als die voor gemaskeerde views. Voor de inhoud van de view maakt dat niet uit.

Alleen de data van de bron wordt beschikbaar gesteld in de ontkoppelviews. Technische velden (zie paragraaf 8.8.3), met uitzondering van de velden m.b.t. de administratieve historie, zullen niet beschikbaar zijn in de ontkoppelviews.

8.5.3 Eén naam, twee schema's

De ontkoppelviews van een bron(module) zullen, per bron(module), in een schema voor gemaskeerde, een schema voor ongemaskeerde, en een schema voor niet aan personen gerelateerde ("overige[GL(11]") ontkoppelviews worden verdeeld:

- Hierdoor kunnen de gemaskeerde en de ongemaskeerde variant van een ontkoppelview dezelfde naam hebben. Het schema geeft aan of het over de gemaskeerde of de ongemaskeerde variant gaat.
  Dit maakt het mogelijk om, binnen de integratiezone/bedrijfszone, dezelfde ETL-processen te gebruiken voor zowel de ongemaskeerde als de gemaskeerde variant van een informatiegebied.
  Zie paragraaf 9.3.3 (Laden van informatiegebieden met twee varianten) voor meer informatie hierover
- Tevens maakt dit het eenvoudiger voor de beheerders om de security met betrekking tot de toegang van data vorm te geven.

N.B. Voor bronnen **zonder persoonsgegevens** zijn twee van de drie schema's (die met views op persoonsgegevens) leeg, voor bronnen met **alleen maar persoonsgegevens** is één schema (dat met views op "overige" gegevens) leeg.

8.5.4 Materialiseren ontkoppelviews en "flip"

De historische ontkoppelviews zullen worden gematerialiseerd. De redenen hiervoor zijn:

- Controle houden op het zichtbaar worden van een data-verversing zodat er geen half-ververste gegevens zichtbaar zijn voor afnemers; de getoonde gegevens zijn altijd, per bronlevering, consistent.
- De beschikbaarheid van de bronzone-data hierin te maximaliseren.

Dit kan door een "flip" te doen van de gematerialiseerde views.
Let op! Voor een gematerialiseerde view is dit normaal niet nodig. We willen dit voor alle tabellen van een bron echter tegelijk doen zodat consistentie[71] van alle binnen het DIM beschikbare gegevens binnen een bron ten allen tijde gegarandeerd kan worden).

[71] Consistentie vanuit replicatie-perspectief; inconsistentie **in** de bron wordt dus keurig gerepliceerd in inconsistentie binnen het DIM.

De actuele ontkoppelviews worden **niet** gematerialiseerd; hun "flip" wordt al gedekt door de flip van de onderliggende historische view.

N.B. Strikt genomen hoeven alleen ontkoppelviews die onderdeel zijn van een gegevensvenster gematerialiseerd te worden; ETL-processen kunnen immers in de actualiteits-metadata controleren of de verversing is afgerond.
We kiezen er toch voor om ontkoppel-views altijd te materialiseren, omdat:
- Dit het beheer vereenvoudigt
- Dit voorkomt dat het creëren van een nieuw informatieproduct (in dit geval een gegevensvenster) impact heeft op de bronzone
- De extra opslagkosten beperkt zijn (zeker bij DXC)

8.6 Master Sequence bronverwerking

Voor elke bronlevering wordt de **master sequence bronverwerking** opgestart, steeds met voor die levering specifieke parameters:
- Bronlevering Naam
- Een standaard-set omgevingsafhankelijke parameters (zie paragraaf 8.6.1)

Starten gebeurt d.m.v. scheduling (door IWS). Deze scheduling is gebaseerd op afspraken in de GLO.

De master sequence start op zijn beurt de ETL-jobs waarin de laadlogica wordt uitgevoerd (meestal harnassen, soms maatwerk).
Bij reruns t.b.v. herstart van een afgebroken verwerking of vanwege een herlevering wordt de master sequence handmatig gestart (door DIM Beheer) waarbij eventueel bepaalde stappen/ETL-jobs daarbinnen (bv. harnassen) overgeslagen kunnen worden.

De master sequence bron-verwerking kent vier standaard-varianten, voor elke leveringstechniek (platte bestanden, export-import, databaselink, berichten) één.

Voor maatwerk kunnen er specifieke master sequences bij komen.

Elk van deze master sequences doorloopt standaard de volgende stappen:
1. Bepaal de voor de staging-harnas benodigde parameters
   (Run ID, Bronlevering ID, Bron Naam)
2. Controleer of vorige levering uit deze bron correct is afgerond
3. Start staging-harnas (verwerking van bronlevering naar staging-laag)
4. Evalueer resultaat staging-harnas
5. Start, bij succes staging-harnas, het bronzone-harnas (verwerking van bronlevering naar staging-laag).
   Geef als parameter ook de, van het staging-harnas ontvangen, procesdatum door.
6. Refresh ontkoppelviews met flip
7. Werk actualiteits-metadata bij

Daarnaast bevat de master sequence huishoudelijke taken (schoning folders e.d.), en eventueel ook "functionele backups" (zie paragraaf 18.3.3).

N.B. Als een bron, naast incrementele bronleveringen, ook periodiek een volledige bronlevering doet (t.b.v. verwijderingen/reconciliatie), dan ontstaat er potentieel een verwerkingsafhankelijkheid tussen die twee leveringen. De aanpak hiervan moet nog bepaald worden[GL(12)].
Zie ook paragraaf 7.2 (Landingszone, hulplocaties en archief)

8.6.1 Standaard-set omgevingsafhankelijke parameters

Op Datastage omgevingsniveau (O, T, A, P) zijn herbruikbare standaard parameter-sets gedefinieerd. Deze worden door de master sequence gebruikt, die ze vervolgens doorgeeft naar de harnassen die door die master sequence aangeroepen worden.

De standaard parameter-set bevat de DIM database connectie gegevens, standaard statische codes/teksten, standaard email gegevens van DIM beheer, en de standaard-definities van de Linux paden op de DataStage-server.

8.7 DLM-tools bronzone

Het Data Lifecycle Management van de bronzone-data wordt deels uitgevoerd door de **laadlogica bronzone**, en deels door de **DLM-tools bronzone**.

> *Deze paragraaf is een placeholder; ontwerp en implementatie van de **DLM-tools bronzone** zijn onderdeel van fase 5. Meer details over deze tools zullen dus in een later stadium aan dit document worden toegevoegd.*

Zie paragraaf 14.3 (DLM in de bronzone~~DLM in de bronzone~~) voor meer informatie.

8.8 Technische aspecten

8.8.1 Database-inrichting

Elke bron (of bronmodule) heeft zowel in de staging-laag als in de bronzone-data een eigen "silo" (in de vorm van een database-schema) voor de opslag van gegevens.

De ontkoppelviews die toegang geven tot deze silo hebben, per paar (met historie, en actueel)[72], elk ook weer een eigen database-schema. De naam van dit schema is gelijk aan dat van de silo, maar met een suffix die aangeeft of het view-paar met gemaskeerde persoonsgegevens, met ongemaskeerde persoonsgegevens, of met overige gegevens betreft.

8.8.2 Toegangsrechten

Alle database-objecten in de bronzone zijn (op A en P) alleen toegankelijk voor geautomatiseerde processen. De bronzone-data is daarnaast toegankelijk voor de ontkoppelviews.

Hierop zijn twee uitzonderingen:

- DBA's kunnen database-objecten wijzigen t.b.v. het aansluiten van nieuwe bronnen of het wijzigen van bestaande leveringen.
  Idealiter loopt dit overigens via een door een release-tool uitgevoerd script, zodat deze toegang niet noodzakelijk is
- DBA/Beheerders/ontwikkelaars kunnen, in geval van een calamiteit, tijdelijk toegang krijgen tot database-objecten t.b.v. onderzoek. Deze tijdelijke toegang verloopt via een "red envelope"-procedure.[73]

[72] Zie paragraaf 8.5.1 (Zes varianten ontkoppelviews)

[73] Zie hoofdstuk 18 (Informatiebeveiliging en –beheer) voor meer informatie over deze procedure.

8.8.3 Technische velden

De hub en de satellieten bevatten een aantal technische velden.

Deze velden zijn hieronder functioneel beschreven; de technische namen worden in de technische documentatie vastgelegd, op basis van de standaarden en richtlijnen op dit gebied.

8.8.3.1 Technische velden hub

Kolom (functionele naam)	Definitie
Hub-sleutel	HASH(SALT + "|" + primaire sleutel) De Hub-sleutel is een tijds-invariante sleutel die een directe unieke relatie heeft met de primaire sleutel van de bron-tabel. Als de primaire sleutel uit meer dan één veld bestaat, dan worden deze velden gescheiden door een pipe-teken. De salt is alleen voor de ETL processen beschikbaar.
Startdatum/tijd	Startdatum/tijd van de versie van het gegeven. Deze is of gebaseerd op de administratieve tijdlijn in de bron of op het extractie moment in de bron.
Einddatum/tijd	Einddatum/tijd van de versie van het gegeven. Deze wordt ook gevuld als er sprake is van een verwijdering in de bron. De einddatum/tijd is gelijk aan de begindatum/tijd van de volgende versie, en daarmee **exclusief**; de versie loopt dus **tot aan** de einddatum/tijd, en **niet tot en met** de einddatum/tijd. De actieve versie heeft altijd einddatum 31-12-9999.
DIM_sleutel hub/satelliet	Een technische DIM-interne sleutel om de hub en de satellieten performant met elkaar te kunnen verbinden. Deze sleutel werkt niet buiten de betreffende hub/satelliet-combinatie.
DIM status	De DIM status geeft aan of de gegevens met deze primaire sleutel fysiek uit de bron verwijderd zijn. Het veld is null als het record nog in de bronlevering zit, en wordt op "zacht verwijderd" gezet indien dat record uit de bron levering verdwenen is. Na de grace periode wordt hij bij de delete op "verwijderd" (en dus niet meer zacht verwijderd)[74] gezet.
Bron	Bevat de bron van het betreffende gegeven. In het geval van gemigreerde data zal hier het betreffende DWH genoemd worden.
DIM datum/tijd aanmaak	Datum/tijd waarop deze versie van het gegeven is aangemaakt binnen het DIM. Representeert dus de administratieve tijdlijn van het DIM zelf. Deze kan afwijken van de administratieve tijdlijn van de bron.
Aangemaakt door	De DataStage-gebruiker welke het gegeven heeft aangemaakt
DIM datum/tijd wijziging	Datum/tijd waarop deze versie van het gegeven is gewijzigd in het DIM, bijvoorbeeld door het toevoegen van de Einddatum/tijd of het invullen van een Archiefvernietiging status. Representeert dus de administratieve tijdlijn van het DIM zelf (niet die van de bron).

[74] Wordt gebruikt om "zacht verwijderen" te kunnen ondersteunen.
Meer informatie hierover in paragraaf 8.3.3 (Volgen van het DLM van de bron: harde en zachte verwijderingen).

Gewijzigd door	De DataStage-gebruiker welke het gegeven heeft gewijzigd
MD5	Een hash die het eenvoudiger maakt om vast te stellen of de data in een nieuwe bronlevering is gewijzigd t.o.v. de bestaande data in het DIM. [75] De hash wordt gedaan over een concatenatie van alle functionele attributen van een tabel opgeslagen in de bronzone (met tussen de velden een separator)
Archiefvernietiging status[76]	Indien de bewaartermijn van een gegeven is overschreden wordt hier de status "te vernietigen" opgevoerd, waarna de controle of het gegeven ook werkelijk vernietigd mag worden kan worden uitgevoerd. Indien dit inderdaad het geval is, dan wordt de status omgezet naar "vernietiging goedgekeurd", is dat niet zo, dan wordt de status omgezet naar "vernietiging uitgesteld".

N.B. Op het moment van schrijven van dit ontwerp bevat de hub nog een redundant veld (DIM bron status). Dit veld zal in een volgende iteratie [GL(13]worden verwijderd.

8.8.3.2 Technische velden satellieten

Kolom (functionele naam)	Definitie
DIM_sleutel hub/satelliet	Een technische DIM-interne sleutel om de hub en de satellieten performant met elkaar te kunnen verbinden. Deze sleutel werkt niet buiten de betreffende hub/satelliet-combinatie.

N.B. Op het moment van schrijven van dit ontwerp bevatten de satellieten nog drie redundante technische velden (Hub-sleutel, Startdatum/tijd en Einddatum/tijd). Deze zullen in een volgende iteratie worden verwijderd.

De technische namen worden in de technische documentatie vastgelegd, op basis van de standaarden en richtlijnen op dit gebied.

8.9 Impact ontwerpuitgangspunten

De ontwerpuitgangspunten zijn meegenomen in (dit deel van) het conceptueel ontwerp. Daarnaast moeten ze meegenomen worden in de detaillering en de realisatie van dat ontwerp.

Ontwerpuitgangspunt	Impact
Bewezen concepten	Bronzone met "single version of the fact" is een "market best practice". Metadata-gedreven harnassen zijn een bewezen concept. De ervaring heeft overigens wel geleerd dat ze bij DataStage-gebaseerde data warehouses weinig worden toegepast.
Geen realtime ambitie	Alle verwerking is batch gedreven

[75] Meer informatie hierover in paragraaf 8.3.2 (Bepaling delta's).

[76] Dit veld wordt alleen in de actieve versie van de hub bijgehouden, en geldt voor alle rijen in de hub/satelliet-combinatie met dezelfde primaire sleutel.

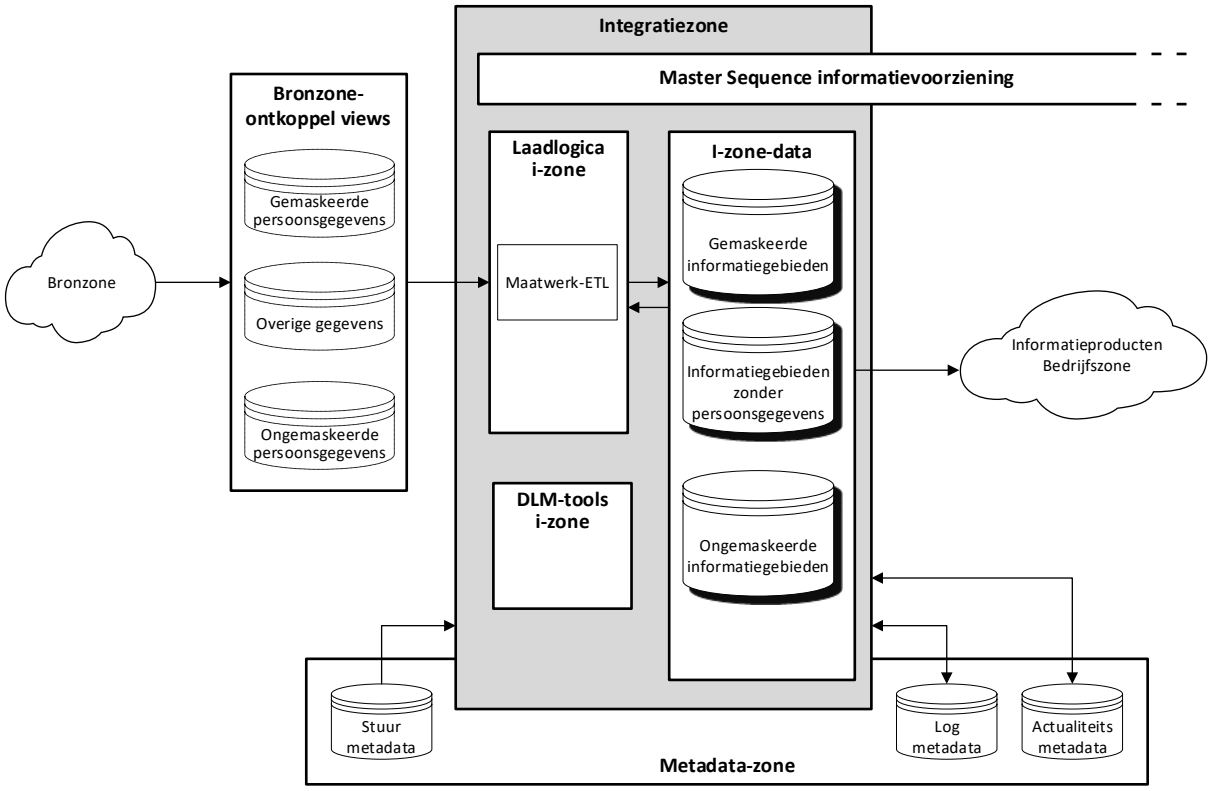
Maximale ontkoppeling	Duidelijke scheiding door middel van ontkoppelviews tussen de bronzone en de bovenliggende lagen. Duidelijke scheiding tussen het verwerken van een bronlevering naar de staging en van de staging naar de bronzone.
Specifiek waar mogelijk, gemeenschappelijk waar noodzakelijk of waardevol	Generieke bouwstenen worden hergebruikt Metadata-gedreven ETL
Kortste keten	De ontkoppelviews verbergen de hub/satelliet-structuren in de bronzone voor de achterliggende zones. Dit maakt de gegevens begrijpelijker, en daardoor vaker direct te gebruiken voor een informatieproduct, bijvoorbeeld door een gegevensvenster direct op de ontkoppelviews van de bronzone te plaatsen.
Compliant	Toegang tot de bronzone alleen via ontkoppelviews (tenzij directe toegang expliciet noodzakelijk bij het afhandelen van calamiteiten). Daardoor strikte scheiding van gemaskeerde en ongemaskeerde gegevens. Toegang voor eindgebruikers altijd via een informatieproduct in de bedrijfszone. Controleprocessen m.b.t. rechtsgrond, proportionaliteit, subsidiariteit (AVG) daardoor eenvoudig in te richten. DLM bronzone conform AVG én Archiefwet.
Eenvoudig	Er worden in de verwerking geen extra stappen gedaan indien er geen functionele noodzaak voor is. De functionaliteit is opgedeeld in logische stappen, die ook als aparte modules gebouwd worden.
Gebruiksvriendelijk	De ontkoppelviews verstoppen de complexiteit van de bronzone.
Gedefinieerd	De data is een 1 op 1 afspiegeling van de administratieve werkelijkheid van de bron. Elk gegeven in de bronzone is helder terug te leiden tot zijn oorsprong. Als afgeleide gegevens noodzakelijk zijn blijven steeds ook de onderliggende gegevens beschikbaar in de bronzone. Een bronlevering wordt óf geheel verwerkt in de bronzone, of (bij technische issues in de levering) helemaal niet; fouten in de brondata zijn immers een deel van de "administratieve werkelijkheid" die, consistent met de bron, ook in het DIM zichtbaar moet blijven. Van elk gegeven zijn definitie en overige relevante metadata vastgelegd in IGC.

Zoveel mogelijk RDBMS-onafhankelijk	Maskering op basis van Optim. Voor overige functionaliteit zoveel mogelijk gebruik maken van DataStage. SQL als tweede keus, geen PL/SQL of andere Oracle-specifieke functies (tenzij gegenereerd door DataStage). Inzet IWS beperkt houden tot opstarten master sequences. Voornaamste afwijking van dit uitgangspunt is het gebruik van gegenereerd SQL (met daarin aanroepen van Optim UDF's binnen Oracle) in de harnassen, i.p.v. een volledig op Datastage gebaseerde oplossing (met daarin aanroepen van masking stages). De rationale hiervoor is: • Een "zuiver" DataStage-harnas moet "op rij-basis" op de ETL-server worden uitgevoerd. Dit bleek een onacceptabel slechte performance te hebben. De nu gebouwde harnassen draaien "op set-basis" en op de RDBMS-server, en leveren daardoor een veel betere performance. • Mocht het DIM verhuizen naar een ander RDBMS, dan heeft dit beperkte impact: alleen de SQL-generatoren in de harnassen moeten worden aangepast. • Dit vereist wel dat dat nieuwe RDBMS ook gebruik kan maken van Optim UDF's (in één of andere vorm). Dit zou een eis moeten zijn bij de keuze van een nieuw RDBMS. Aandachtspunt bij het inrichten van de maskering m.b.v. UDF's is dat deze maskering functioneel identiek moet kunnen worden uitgevoerd in toekomstige non-RDBMS onderdelen van het DIM. Een m.b.v. een UDF gemaskeerde BSN, bijvoorbeeld, moet dezelfde waarde opleveren als een binnen de datalake-zone (met Optim, maar niet m.b.v. een UDF) gemaskeerde BSN. Een tweede afwijking is het gebruik van materialized views voor de ontkoppelviews. De rationale hiervoor is: • Het (bijna) 1:1 kopiëren van gegevens is veel eenvoudiger en performanter in te richten met materialized views dan met ETL-jobs. • Mocht het DIM verhuizen naar een ander RDBMS, dan heeft dit beperkte impact: materialized views (of equivalenten daarvan) zijn voor de meeste RDBMS'en beschikbaar, en het migreren zelf is, door de generieke opzet van de ontkoppelviews, eenvoudig.
Lineage mag niet gebroken worden	Op dit moment start de horizontale lineage bij de ontkoppelviews. De ontkoppelviews mogen daarom geen logica bevatten (anders dan het verbergen van de hub/satelliet-structuren). Hierdoor kunnen ze worden beschouwd als een ongewijzigde afspiegeling van (de historie van) de door de bronnen geleverde gegevens.

Minimale, en beheerbare, toolset	Oplossingen gebaseerd op DataStage en Optim, tenzij duidelijk te argumenteren valt dat het daarmee niet op te lossen valt. Daarnaast bij de afdeling DWH reeds bekende (en door UWV toegestane) tools. Inzet IWS alleen voor opstarten master sequences, zodat de hiervoor benodigde kennis beperkt is.

9 INTEGRATIEZONE

9.1 Basisopzet



In de **integratiezone** worden afgeleide en/of geïntegreerde gegevens opgeslagen die, om redenen van efficiëntie of consistentie, van belang (of verplicht) zijn voor meerdere **informatieproducten** in de **bedrijfszone**.

De **integratiezone** bestaat uit **informatiegebieden**. Elk van deze informatiegebieden is gebaseerd op gegevens uit de bronzone (van één of meer bronnen), al dan niet gecombineerd met gegevens uit andere informatiegebieden in de integratiezone, en meestal bedoeld voor een specifieke groep informatieproducten.

N.B. In de migratiestrategie en de deliverables van het project DataFabriek is ook sprake van informatiegebieden. De term wordt daar gebruikt voor, vanwege hun grote samenhang, in één keer te migreren groepen informatieproducten, al dan niet (deels) gebaseerd op gegevens uit de integratiezone. Gepoogd zal worden om in de migratiestrategie en deliverables hiervoor de term "migratieblok" te introduceren, om misverstanden in de toekomst zo te beperken.

De integratiezone is "**vraaggedreven**"; de rechtvaardiging voor een informatiegebied volgt altijd uit eisen vanuit de erop te baseren informatieproducten (nu of in de nabije toekomst).[77] De integratiezone is dus **niet** bedoeld om alle bronzone-data in één geïntegreerd model te gieten.

[77] Deze opzet is vergelijkbaar met die van UGL (de laag in DWH 3.0 waar de halffabrikaten in staan)
Vandaar dat deze laag ook als startpunt wordt gebruikt bij het ontwerp van de diverse DIM-informatiegebieden; de UGL-modellen zullen worden beoordeeld en waar mogelijk/zinnig zoveel mogelijk overgenomen worden. Uiteraard wordt e.e.a. nog wel aangepast om te passen binnen de totaal-opzet van het DIM, en om recht te doen aan eventuele geleerde lessen bij DWH 3.0.

Daarnaast geldt dat een informatieproduct in de bedrijfszone **niet per definitie** gebaseerd is op een informatiegebied in de integratiezone. Als dat informatieproduct direct op de bronzone-data (eigenlijk: de bronzone-ontkoppelviews) kan worden gebaseerd heeft dat de voorkeur. [78]

In een informatiegebied worden gemaskeerde en ongemaskeerde gegevens nooit gecombineerd.

Er zijn daarom drie typen informatiegebieden:
- Informatiegebieden met gemaskeerde persoonsgegevens ("Gemaskeerde informatiegebieden")
- Informatiegebieden met ongemaskeerde persoonsgegevens ("Ongemaskeerde informatiegebieden")
- Informatiegebieden zonder persoonsgegevens

Soms bestaat er voor een informatiegebied een gemaskeerde en een ongemaskeerde variant; gelijk qua structuur en afleidingen, en met als enig verschil dat de eerste gebaseerd is op gemaskeerde bronzone-data, en de tweede op ongemaskeerde.

De **laadlogica integratiezone** transformeert de gegevens uit de bronzone-ontkoppelviews (al dan niet gecombineerd met gegevens uit andere informatiegebieden in de integratiezone) en slaat deze op in een informatiegebied.

De **laadlogica integratiezone** is, anders dan bij de laadlogica t.b.v. **staging-laag** en **bronzone-data**, maar beperkt te standaardiseren. Er wordt daarom geen gebruik gemaakt van harnassen; **maatwerk-ETL** is de standaard.
Deels wordt deze laadlogica overigens nog steeds gedreven door **stuur-metadata**.

Binnen het maatwerk-ETL wordt zoveel mogelijk gebruik gemaakt van standaard-bouwstenen, bijvoorbeeld t.b.v. logging.

De **laadlogica integratiezone** wordt opgestart door de **master sequence informatievoorziening**. Dit is een overkoepelende aansturingscomponent die zowel de **laadlogica integratiezone** (in deze zone) als het aanmaken van de erop gebaseerde informatieproducten (in de **bedrijfszone**) aanstuurt.

De **master sequence informatievoorziening** wordt **niet** doorgestart vanuit de **laadlogica bronzone** en/of de **master sequence bronverwerking**, maar draait op een eigen (geschedulede) hartslag.

De **laadlogica integratiezone** controleert vervolgens zelf, na gestart te zijn, in de door de bronzone geleverde **actualiteits-metadata** of aan alle randvoorwaarden voor het verversen van een informatiegebied is voldaan. Is dat niet zo dan blijft de laadlogica wachten tot dat wel zo is.

De verwerkingsresultaten van de **laadlogica integratiezone** (succes, ondervonden fouten, etc) komen in de **log-metadata**.
Na het verversen van een informatiegebied wordt ook in de integratiezone de **actualiteits-metadata** bijgewerkt, zodat (m.n. voor afnemende ETL-processen) eenvoudig te achterhalen is tot op welk punt in de tijd een de gegevens in een informatiegebied zijn bijgewerkt.

Het Data Lifecycle Management van de gegevens in de diverse informatiegebieden wordt deels uitgevoerd door de **laadlogica integratiezone**, en deels door de **DLM-tools integratiezone**.

[78] E.e.a. conform het "kortste keten" principe

Anders dan bij de bronzone maakt de integratiezone géén gebruik van ontkoppelviews bij het ter beschikking stellen van de informatiegebieden. Ontkoppelviews zijn niet nodig, omdat:

1. De technische structuur van de informatiegebieden al geoptimaliseerd is voor gebruik in informatieproducten;
2. Gemaskeerde en ongemaskeerde persoonsgegevens nooit in één informatiegebied kunnen voorkomen;
3. De informatiezone geen "zachte verwijderingen" toepast om met onvolledige bronleveringen om te kunnen gaan (zoals in de bronzone)..
   Logische verwijderingen, zoals standaard bij "slowly changing dimensions", komen uiteraard wel voor.
4. Het ontkoppelen van verversingen, indien nodig, binnen de bedrijfszone plaatsvindt.

*Ontwerp en implementatie van de **DLM-tools integratiezone** zijn onderdeel van fase 5. Meer details over deze tools zullen dus in een later stadium aan dit document worden toegevoegd.*

9.2 Informatiegebieden

9.2.1 Scoping van een informatiegebied

9.2.1.1 Logisch samenhangend

Een informatiegebied bevat een logisch samenhangende groep afgeleide en/of geïntegreerde gegevens, gebaseerd op één of meer bronnen, meestal gebruikt om de afleidingen/integraties voor een, ook logisch samenhangende groep informatieproducten centraal uit te voeren en op te slaan.

Ook hermodelleren (bijvoorbeeld het creëren van conforme dimensies t.b.v. meerdere datamarts) en verrijken (bijvoorbeeld defaulting, koppelen met referentiegegevens) valt onder dit afleiden/integreren.

De rationale hiervoor kan zijn:

- **Consistentie**
  Garantie dat alle achterliggende informatieproducten dezelfde (afgeleide) gegevens gebruiken.
- **Efficiëntie**
  Maar één keer uitvoeren van een afleiding, i.p.v. voor elk informatieproduct opnieuw.

N.B. Als de gegevens in een informatiegebied al in ongewijzigde vorm (dus zoals beschikbaar in de ontkoppelviews op de bronzone) bruikbaar zouden zijn in de informatieproducten waar het informatiegebied voor bedoeld is, dan hoeven deze gegevens **niet** in het informatiegebied te worden opgenomen; conform het "kortste keten" principe kunnen die informatieproducten zich dan baseren op een combinatie van de bronzone voor de "ruwe" brondata en de integratiezone voor afgeleide data.
Alleen als dit tot onoverkomelijke bezwaren in de historieafhandeling leidt wordt hiervan afgeweken.

9.2.1.2 Gegevenstechnisch samenhangend

Als de logische samenhang (zie vorige paragraaf) leidt tot een informatiegebied met, binnen dat informatiegebied, zeer ongelijksoortige vereisten aan de laadlogica en/of sterk afwijkende afhankelijkheden met de bron(nen), dan is het vaak handig om dat informatiegebied verder op te knippen.

Opknippen is ook handig als de bewaartermijnen voor de gegevens binnen het informatiegebied sterk verschillen.

9.2.1.3 Bedoeld voor meerdere informatieproducten

Een informatiegebied is meestal bedoeld voor een specifieke groep informatieproducten.

Is er maar één afnemend informatieproduct (en blijft dat ook zo), dan is het in het algemeen beter om géén informatiegebied in te richten, maar de afleidingen/integraties onderdeel te maken van het informatieproduct zelf ("kortste keten" principe).

Een bijzonder geval zijn informatiegebieden die brongegevens bruikbaar maken voor **elk** verder gebruik binnen het DIM. Dit is bijvoorbeeld het geval als een bron gegevensmutaties aanlevert aan het DIM (insert, update, delete). Voor dergelijke bronnen worden deze mutaties "as is" in de bronzone opgeslagen, en vervolgens in een informatiegebied omgevormd tot een gegevenshistorie langs één of twee tijdlijnen (afhankelijk van de behoefte van de DIM-afnemers).

N.B. Een dergelijk informatiegebied zal dus vaak als input dienen voor andere informatiegebieden.

9.2.1.4 Standaard in volledig detail

Standaard worden gegevens in volledig detail in een informatiegebied opgenomen. Het informatiegebied moet immers meerdere informatieproducten, potentieel met verschillende aggregatie-eisen, kunnen ondersteunen.

Dat betekent ook dat de inputgegevens alleen gefilterd worden (om functionele redenen of vanwege een te lage datakwaliteit) als dat filter voor alle achterliggende informatieproducten relevant is.
Zelfs dan verdient het in veel gevallen de voorkeur om een functioneel filtercriterium of een kwaliteitsindicatie als afgeleid veld in het informatiegebied op te nemen, zodat per informatieproduct kan worden bepaald of het filter moet worden toegepast (en het toepassen van het filter daar tegelijkertijd heel eenvoudig is).

9.2.1.5 Zo simpel mogelijk

Aangezien een informatiegebied altijd weer opnieuw uit de bronzone kan worden opgebouwd hoeven transformaties, consolidaties en/of standaardisaties die geen (verwachte) waarde hebben in de achterliggende informatieproducten niet "voor de zekerheid en/of volledigheid" alvast te worden ingericht.
Aandachtspunt daarbij is wel dat de structuur van een informatiegebied, vanwege de impact op reeds erop gebaseerde informatieproducten, niet eenvoudig achteraf wijzigbaar is.
Deze structuur dient dus wel ~~sig~~alleen de gegevensbehoefte van reeds bekende informatieproducten

9.2.1.6 Gemaskeerd en ongemaskeerd gescheiden

Een informatiegebied mag nooit zowel gemaskeerde als ongemaskeerde persoonsgegevens bevatten. Lijkt hiervoor toch een noodzaak te bestaan, dan is dat waarschijnlijk te wijten aan een non-compliant afnemerseisen en/of een verkeerd ontworpen informatieproduct (in de bedrijfszone).

Soms bestaat er voor een informatiegebied wél een gemaskeerde en een ongemaskeerde variant; gelijk qua structuur en afleidingen, en met als enig verschil dat de eerste gebaseerd is op gemaskeerde bronzone-data, en de tweede op ongemaskeerde.

9.2.2 Gemaskeerde en ongemaskeerde informatiegebieden

In een informatiegebied worden gemaskeerde en ongemaskeerde gegevens nooit gecombineerd.

Er zijn daarom drie typen informatiegebieden te onderkennen:

- Informatiegebieden met gemaskeerde persoonsgegevens ("Gemaskeerde informatiegebieden")
- Informatiegebieden met ongemaskeerde persoonsgegevens ("Ongemaskeerde informatiegebieden")
- Informatiegebieden zonder persoonsgegevens

9.2.2.1 Gemaskeerde informatiegebieden

Deze informatiegebieden zijn gebaseerd op gemaskeerde bronzone-gegevens en/of andere gemaskeerde informatiegebieden, al dan niet gecombineerd met "overige gegevens" uit de bronzone en/of gegevens uit "informatiegebieden zonder persoonsgegevens".

N.B. Ongemaskeerde bronzone-gegevens en gegevens uit ongemaskeerde informatiegebieden mogen **niet** als input voor een gemaskeerd informatiegebied worden gebruikt.

9.2.2.2 Ongemaskeerde informatiegebieden

Deze informatiegebieden zijn gebaseerd op ongemaskeerde bronzone-gegevens en/of andere ongemaskeerde informatiegebieden, al dan niet gecombineerd met "overige gegevens" uit de bronzone en/of gegevens uit "informatiegebieden zonder persoonsgegevens".

N.B. Gemaskeerde bronzone-gegevens en gegevens uit gemaskeerde informatiegebieden mogen **niet** als input voor een ongemaskeerd informatiegebied worden gebruikt.

9.2.2.3 Informatiegebieden zonder persoonsgegevens

Deze informatiegebieden zijn óf uitsluitend gebaseerd op "overige gegevens" en/of gegevens uit "informatiegebieden zonder persoonsgegevens", óf bevatten aggregaten van persoonsgegevens (gemaskeerd of ongemaskeerd) die door die aggregatie anoniem zijn geworden (en dus niet meer als persoonsgegeven beschouwd hoeven te worden).

In dat tweede geval kunnen die aggregaten weer gecombineerd zijn met "overige gegevens" uit de bronzone en/of gegevens uit "informatiegebieden zonder persoonsgegevens".

N.B. Gemaskeerde en ongemaskeerde bronzone-gegevens, en gegevens uit gemaskeerde en ongemaskeerde informatiegebieden, mogen **alleen** als input voor een informatiegebied zonder persoonsgegevens worden gebruikt als ze bij het laden in dat informatiegebied door aggregatie anoniem worden gemaakt.

Het verdient de voorkeur om "anonieme" aggregaten van persoonsgegevens te baseren op gemaskeerde gegevens, zodat het DLM consistent blijft met andere voor rapportage/analyse bedoelde gegevens.

In twee gevallen kan hiervan afgeweken worden:

1. Het aggregaat is, in informatieproducten, alleen relevant in combinatie met ongemaskeerde gegevens.
   In dat geval wordt het aggregaat, om consistent te blijven met de detailgegevens, in het algemeen ook gebaseerd op ongemaskeerde gegevens.
2. Het aggregaat wordt gebruikt in informatiegebieden met twee varianten (zie volgende paragraaf).
   In dat geval wordt het aggregaat ook in twee varianten aangemaakt.

N.B. In alle gevallen (dus ook bij de standaardsituatie) dient in de precieze definitie van het aggregaat (zoals opgeslagen bij de "functionele termen" in de metadata) ook opgenomen te

zijn op welke gegevens (gemaskeerd of ongemaskeerd) het aggregaat is gebaseerd, en dus welk DLM op het aggregaat van toepassing is.

9.2.2.4 Informatiegebieden met twee varianten

Soms bestaat er van een informatiegebied een gemaskeerde en een ongemaskeerde variant; gelijk qua structuur en afleidingen, en met als enig verschil dat de eerste gebaseerd is op gemaskeerde bronzone-data, en de tweede op ongemaskeerde.

Dit zal vooral voorkomen als er in de bedrijfszone op vergelijkbare gegevens zowel gemaskeerde informatieproducten (t.b.v. analyse/rapportage) als ongemaskeerde informatieproducten (t.b.v. operationeel gebruik) zijn gedefinieerd.

Deze twee varianten worden ververst met dezelfde laadlogica; het enige verschil is de te gebruiken input-gegevens (gemaskeerd of ongemaskeerd) uit bronzone en/of andere informatiegebieden, en deze keuze wordt middels stuur-metadata aan de laadlogica meegegeven.

Governance-technisch blijven het twee informatiegebieden; doelproces en doelgroep van de achterliggende informatieproducten zijn immers totaal verschillend.

N.B. Bij informatiegebieden met twee varianten zullen eventuele bijbehorende aggregaten (zie vorige paragraaf) meestal ook in twee varianten moeten worden aangemaakt.
In afwijking van de standaard zullen deze aggregaten **niet** in een "informatiegebied zonder persoonsgegevens" worden opgenomen.
In plaats daarvan worden de op ongemaskeerde gegevens gebaseerde aggregaten onderdeel van een informatiegebied met ongemaskeerde persoonsgegevens, en de op gemaskeerde gegevens gebaseerde aggregaten onderdeel van een informatiegebied met gemaskeerde persoonsgegevens.
Dit om verdere verwerking in (de twee versies van) de informatieproducten ook zo gelijk mogelijk te houden.

9.2.3 **DLM integratiezone t.o.v. de bronzone**

De bronzone dient als stabiele databasis. Hierdoor kan, binnen de bewaartermijnen van die bronzone, een informatiegebied binnen de integratiezone altijd weer opnieuw opgebouwd worden.

In principe zal een informatiegebied bij verversen op een slimme manier opnieuw worden opgebouwd tenzij er een functionele vereiste is om historische rapportageresultaten constant te houden, en deze vereiste niet kan worden gedekt met een versionering o.b.v. de administratieve tijdlijn.
Zie paragraaf 14.4 (DLM in de integratiezone~~DLM in de integratiezone~~) voor een gedetailleerde beschrijving van het DLM binnen de integratiezone.

9.2.4 **Datamodel**

9.2.4.1 Eigenaarschap

Van alle gegevenselementen in de **integratiezone** dient duidelijk te zijn wat deze gegevenselementen betekenen en hoe deze gegevenselementen gevuld dienen te worden:

- Ieder gegevenselement in de **integratiezone** moet gekoppeld zijn aan een functionele term afkomstig uit de functionele terminologie laag in IGC. Deze koppeling zal gemaakt worden in het datamodel van de integratiezone in IDA en raadpleegbaar zijn voor de eindgebruiker in IGC. Bij voorkeur dient de functionele term afkomstig te zijn uit een FUGEM, CGM of HUDSV. Als dit niet mogelijk is, bijvoorbeeld wanneer sprake is van een

niet op HUDSV gebaseerd afgeleid veld, dan wordt een functionele term toegevoegd aan IGC.
De definitie van een functionele term en de eigenaar van een functionele term worden vastgelegd in IGC.

- Voor ieder gegevenselement in de **integratiezone** moet duidelijk gemaakt worden hoe deze kolom gevuld dient te worden. Dit zal veelal vanuit gegevenselementen van de **bronzone** zijn. De mapping wordt in IDA vastgelegd en zal, als onderdeel van het Functioneel Ontwerp, als basis dienen voor het ontwikkelwerk van de technische ontwikkelteams. Het resultaat zal voor de eindgebruikers raadpleegbaar zijn in IGC.

9.2.4.2 Gebruikte modellerings-concepten

De modellering van een informatiegebied is afhankelijk van de erdoor te ondersteunen informatieproducten.

Zowel modellering in 3NF (plus historie) als dimensionele structuren (m.n. conforme dimensies voor groepen datamarts) zullen voorkomen.
In theorie zijn ook "platte" (gedenormaliseerde) structuren mogelijk.

De modelleeroplossing zal enerzijds specifiek gemaakt zijn voor het correcte gebruik binnen een informatiegebied, en anderzijds generiek gemaakt zijn om hergebruik voor andere informatiegebieden mogelijk te maken.

9.2.4.3 Versionering

Ook de versionering binnen een informatiegebied is afhankelijk van de erdoor te ondersteunen informatieproducten.

Vaak zal dit versionering langs twee tijdlijnen (administratief en geldigheid) zijn. Dit om reproduceerbaarheid van S&V-rapportages te kunnen ondersteunen.

Versionering enkel op geldigheid (dus gebaseerd op alleen de administratief actieve gegevens uit de bronzone), of zelfs geen enkele versionering (alleen administratief én qua geldigheid actieve gegevens) is mogelijk.

Voor informatiegebieden die tijd-gedreven informatieproducten[79] moeten kunnen ondersteunen is versionering langs de administratieve tijdlijn verplicht!

N.B. In alle gevallen dient in de precieze definities van (de gegevenselementen in) het informatiegebied (zoals opgeslagen bij de "functionele termen" in de metadata) ook opgenomen te zijn hoe de gegevens geversioneerd zijn.

[79] Zie paragraaf 10.3.3 (Tijd-gedreven of gegevens-gedreven verversing)

9.3 Laadlogica integratiezone

9.3.1 Maatwerk, maar wel met een standaard-aanpak

Hoewel de detailstappen voor elk informatiegebied verschillen zijn er ook een aantal stappen en stukken functionaliteit die voor alle informatiegebieden gelijk zijn.

De altijd te doorlopen stappen zijn:

1 Start
Haal, m.b.v. de door de master sequence meegegeven parameters, de voor het informatiegebied benodigde stuur-metadata op.

Maak een record aan in de log-metadata met als status "lopend".

2 Initialisatie
Bepaal, op basis van DataStage omgevingsvariabelen, de omgevingsafhankelijke[80] delen van paden en eventuele bestandsfolders.

3 Detecteer beschikbaarheid input-gegevens
Op basis van de actualiteits-metadata.

4 Ververs het informatiegebied.

X1 Verzend notificaties
Meld succes of issues via e-mail aan de relevante partijen.
Welke meldingen naar welke partijen moeten, en welke emailadressen voor deze partijen gebruikt moeten worden, staat in de stuur-metadata.

X2 Einde job

Vervang, in de log-metadata, de markering "lopend" door "klaar" of "afgebroken", afhankelijk van het resultaat van de verwerking tot op dit punt.
Werk de actualiteits-metadata bij.

[einde ETL-job]

Binnen het **maatwerk-ETL** wordt zoveel mogelijk gebruik gemaakt van steeds hetzelfde stappenplan en steeds dezelfde standaard-bouwstenen (bijvoorbeeld t.b.v. logging).
Dit stappenplan wordt echter waarschijnlijk niet standaard afgedwongen door een harnas of ander ETL-raamwerk.

9.3.2 Verversing altijd gegevens-gedreven

Informatieproducten (in de bedrijfszone) kunnen tijd-gedreven of gegevens-gedreven ververst moeten worden (meer details in paragraaf 10.3.3,Tijd-gedreven of gegevens-gedreven verversing).
Verversing van de informatiegebieden is echter altijd gegevens-gedreven, en start zodra de input-gegevens voor het informatiegebied beschikbaar zijn.
Dit om te voorkomen dat er complexe tijds-afhankelijkheden ontstaan tussen de verversing van informatiegebieden en die van informatieproducten.

[80] Met "omgevingsafhankelijk" wordt bedoeld: afhankelijk van de OTAP-omgeving waarin het proces draait.

De laadlogica voor een informatiegebied controleert, na door de master sequence informatievoorziening gestart te zijn, dus altijd in de stap "Detecteer beschikbaarheid input-gegevens" in de actualiteits-metadata of alle vereiste gegevens ververst zijn, en blijft, indien nodig, deze metadata "pollen" tot zulks het geval is.

9.3.3 Laden van informatiegebieden met twee varianten

De ETL-job voor het vullen van een gemaskeerde of ongemaskerde variant van een informatiegebied zal dezelfde zijn. Dit kan doordat de input-gegevens en de output-gegevens qua gegevensstructuur en tabel/view/kolom-namen identiek zijn voor zowel het gemaskeerde als het ongemaskeerde informatiegebied. Het verschil zit in de schema-namen. Hierdoor is het mogelijk deze als parameters voor de ETL-job te gebruiken.

9.4 Master sequence informatievoorziening

Net als het verwerken van bronleveringen zal er ook het verversen van informatiegebieden en het verversen/leveren van informatieproducten worden bestuurd met master sequences, die zelf weer, op de hartslag van het DIM, door IWS worden gescheduled.

Ontwerp van deze master sequence valt in fase 5, en zal dan in dit ontwerp worden opgenomen.

9.5 DLM-tools integratiezone

Het Data Lifecycle Management van de informatiegebieden in de integratiezone wordt deels uitgevoerd door de **laadlogica bronzone**, en deels door de **DLM-tools integratiezone**.

*Deze paragraaf is een placeholder; ontwerp en implementatie van de **DLM-tools integratiezone** zijn onderdeel van fase 5. Meer details over deze tools zullen dus in een later stadium aan dit document worden toegevoegd.*

Zie paragraaf 14.4 (DLM in de integratiezone~~DLM in de integratiezone~~) voor meer informatie.

9.6 Technische aspecten

9.6.1 Database-inrichting

Elk informatiegebied heeft een eigen databaseschema om het toegangsbeheer en de beveiliging te vergemakkelijken.

Elke schemanaam heeft een suffix die aangeeft of het een informatiegebied met gemaskeerde persoonsgegevens, met ongemaskeerde persoonsgegevens, of met overige gegevens (geen persoonsgegevens) betreft.
Voor een informatiegebied met twee varianten zijn alle tabel- en kolomnamen voor deze twee varianten gelijk, en verschilt in de schemanaam alleen de suffix.

9.6.2 Toegangsrechten

Alle database-objecten in de integratiezone zijn (op A en P) alleen toegankelijk voor geautomatiseerde processen. De informatiegebieden zijn daarnaast toegankelijk voor gegevensvensters.

Hierop zijn twee uitzonderingen:

- DBA's kunnen database-objecten wijzigen t.b.v. het creëren van nieuwe of het wijzigen van bestaande informatiegebieden.
  Idealiter loopt dit overigens via een door een release-tool uitgevoerd script, zodat deze toegang niet noodzakelijk is
- DBA/Beheerders/ontwikkelaars kunnen, in geval van een calamiteit, tijdelijk toegang krijgen tot database-objecten t.b.v. onderzoek. Deze tijdelijke toegang verloopt via een "red envelope"-procedure.[81]

9.6.3 Technische velden

Niet alle tabellen in de informatiegebieden bevatten dezelfde technische velden. Dit omdat de modellering/versionering per informatiegebied kan verschillen.

Technische velden met dezelfde functie zullen echter wel altijd, in alle informatiegebieden, dezelfde naam en afleiding hebben.

Kolom (functionele naam)	Definitie
Startdatum/tijd	Startdatum/tijd van de administratieve versie van het gegeven. *(Alleen relevant voor tabellen met een versionering langs de administratieve tijdlijn)*
Einddatum/tijd	Einddatum/tijd van de administratieve versie van het gegeven. Deze wordt ook gevuld als er sprake is van een logische verwijdering. De einddatum/tijd is gelijk aan de begindatum/tijd van de volgende versie, en daarmee **exclusief**; de versie loopt dus **tot aan** de einddatum/tijd, en **niet tot en met** de einddatum/tijd. De actieve versie heeft altijd einddatum 31-12-9999. *(Alleen relevant voor tabellen met een versionering langs de administratieve tijdlijn)*
DIM-sleutel	Een technische, binnen het DIM gegenereerde sleutel die als primaire sleutel voor de tabel fungeert. *(binnen de integratiezone beperkt relevant; om parallel laden te kunnen ondersteunen kan deze DIM-sleutel geen bredere relevantie hebben dan het informatiegebied zelf, plus eventueel daar weer op gebaseerde informatiegebieden)*
DIM datum/tijd aanmaak	Datum/tijd waarop deze versie van het gegeven is aangemaakt binnen de integratiezone. Dit kan afwijken van de administratieve tijdlijn van de bron.
Aangemaakt door	De DataStage-gebruiker welke het gegeven heeft aangemaakt

[81] Zie hoofdstuk 18 (Informatiebeveiliging en –beheer) voor meer informatie over deze procedure.

DIM datum/tijd wijziging	Datum/tijd waarop deze versie van het gegeven is gewijzigd in het DIM, bijvoorbeeld door het toevoegen van de Einddatum/tijd of het invullen van een Archiefvernietiging status.
Gewijzigd door	De DataStage-gebruiker welke het gegeven heeft gewijzigd
Archiefvernietiging status[82]	Indien de bewaartermijn van een gegeven is overschreden wordt hier de status "te vernietigen" opgevoerd, waarna de controle of het gegeven ook werkelijk vernietigd mag worden kan worden uitgevoerd. Indien dit inderdaad het geval is, dan wordt de status omgezet naar "vernietiging goedgekeurd", is dat niet zo, dan wordt de status omgezet naar "vernietiging uitgesteld".

De technische namen worden in de technische documentatie vastgelegd, op basis van de standaarden en richtlijnen op dit gebied.

9.7 Impact ontwerpuitgangspunten

De ontwerpuitgangspunten zijn meegenomen in (dit deel van) het conceptueel ontwerp. Daarnaast moeten ze meegenomen worden in de detaillering en de realisatie van dat ontwerp.

Ontwerpuitgangspunt	**Impact**
Bewezen concepten	Een verplichte integratiezone in één organisatie-breed datamodel ("single version of the truth") is, als concept, bij de meeste DWH-implementaties verlaten. Nodeloos complex, en met weinig toegevoegde waarde. Een modulaire,"vraaggedreven" integratiezone, daarentegen, is een "market best practice", die zich ook al bij DWH3 heeft bewezen.
Geen realtime ambitie	Alle verwerking is batch gedreven
Maximale ontkoppeling	Het verversen van informatiegebieden is niet hard gekoppeld aan de levering van de vereiste bronnen, maar gescheduled op DIM-hartslag, met vervolgens een eigen controle op de beschikbaarheid van de benodigde bronzone-gegevens.
Specifiek waar mogelijk, gemeenschappelijk waar noodzakelijk of waardevol	Generieke bouwstenen worden hergebruikt Sturing middels standaard master sequences.
Kortste keten	Geen informatiegebieden implementeren als de informatieproducten ook met bronzone-data toekunnen.

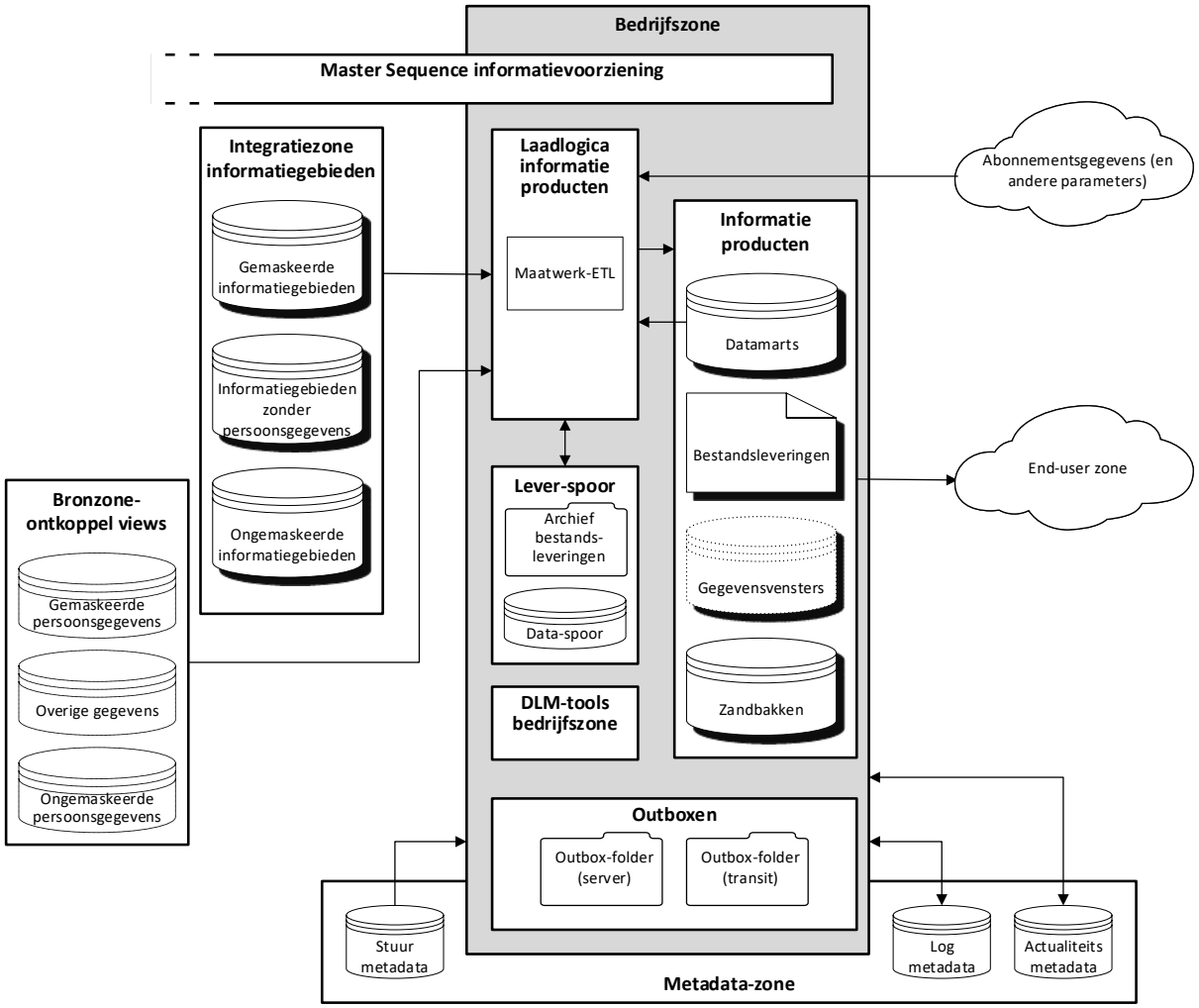
[82] Dit veld wordt alleen in de actieve versie van de tabel bijgehouden, en geldt voor alle rijen in de tabel met dezelfde primaire sleutel.

Compliant	Strikte scheiding van gemaskeerde en ongemaskeerde gegevens; een informatiegebied mag maar één van beide bevatten. Toegang voor eindgebruikers altijd via een informatieproduct in de bedrijfszone. Controleprocessen m.b.t. rechtsgrond, proportionaliteit, subsidiariteit (AVG) daardoor eenvoudig in te richten. DLM integratiezone conform AVG én Archiefwet. Voldoende rijk (qua attributen) om rij-filters (naar of in informatieproduct) eenvoudig te kunnen implementeren.
Eenvoudig	Er worden in de verwerking geen extra stappen gedaan indien er geen functionele noodzaak voor is. De functionaliteit is opgedeeld in logische stappen, met optimaal gebruik van herbruikbare bouwstenen. Geen transformaties, consolidaties en/of standaardisaties die geen (verwachte) waarde hebben in de achterliggende informatieproducten.
Gebruiksvriendelijk	Datamodel en data eenvoudig te gebruiken in achterliggende informatieproducten (bv. datamarts), zowel qua structuur als qua "voorberekende" afleidingen. Datamodel zo leesbaar/begrijpelijk mogelijk (binnen eventuele compliance-beperkingen), zowel voor de Datafabriek zelf, als, bij self service, door de afnemer. Datamodellering niet generiek en/of gestandaardiseerd, maar geoptimaliseerd voor gebruik binnen een informatieproduct.

Gedefinieerd	Eigenaarschap definities/afleidingen belegd (bij bron, Gegevensdiensten of afnemer) en vastgelegd. Afleidingen gevalideerd/geaccordeerd door eigenaar, en gedocumenteerd. Alleen vereiste (kwaliteits)filters, en deze altijd gedocumenteerd/gevalideerd/geaccordeerd. Horizontale lineage vanaf de bronzone-ontkoppelviews. Verticale lineage door verwerking relevante delen FO in IDA/IGC.
Zoveel mogelijk RDBMS-onafhankelijk	Zoveel mogelijk gebruik maken van DataStage. SQL als tweede keus, geen PL/SQL of andere Oracle-specifieke functies (tenzij gegenereerd door DataStage).
Lineage mag niet gebroken worden	Alleen op DataStage gebaseerde transformaties tussen bronzone-ontkoppelviews en informatiegebieden. Daardoor automatisch gegeneerde horizontale lineage.
Minimale, en beheerbare, toolset	Oplossingen gebaseerd op DataStage, tenzij duidelijk te argumenteren valt dat het daarmee niet op te lossen valt. Daarnaast bij de afdeling DWH reeds bekende (en door UWV toegestane) tools. Inzet IWS alleen voor opstarten master sequences, zodat de hiervoor benodigde kennis beperkt is.

10 BEDRIJFSZONE

10.1 Basisopzet



In de **bedrijfszone** worden **informatieproducten** ter beschikking gesteld aan afnemers, t.b.v. gebruik in de **end-user zone**.

Er zijn vier typen **informatieproducten**:

- **Datamarts**
  Op maat gemaakte gegevensverzamelingen, vaak dimensioneel gestructureerd, ten behoeve van specifieke gebruikersgroepen en/of -processen (meestal rapportages, OLAP, dashboards, of vergelijkbaar), inclusief self service BI. Vaak vereist het op maat maken complexe transformaties en afleidingen.
- **Gegevensvensters**
  Virtuele gegevensverzamelingen (deelverzamelingen van de in het DIM beschikbare gegevens) die ontsloten worden ten behoeve van self-service BI en/of self service analytics, d.w.z. rechtstreeks gebruik door afnemers, bijvoorbeeld t.b.v. querying en ad-hoc rapportage. Anders dan bij datamarts worden er binnen een gegevensvenster geen complexe transformaties en/of afleidingen **uitgevoerd**. De logica beperkt zich tot filters op tabel-, rij- en attribuutniveau, om te garanderen dat de getoonde deelverzameling alleen gegevens bevat waarvoor rechtsgrond geldt, en waarvan het gebruik voldoet aan de AVG-principes m.b.t. proportionaliteit en subsidiariteit.

Het gegevensvenster kan overigens wel complexe afleidingen bevatten, bijvoorbeeld omdat die onderdeel zijn van een door het gegevensvenster ontsloten informatiegebied.

- **Bestandsleveringen**[83]
  Levering van in meer of mindere mate bewerkte gegevens, doorgaans in de vorm van platte bestanden. Voorbeelden zijn leveringen aan pensioenfondsen en aan analyse-omgevingen.
- **Zandbakken**
  Data-omgevingen t.b.v. analyse.
  Een zandbak kan virtueel of fysiek zijn, of een combinatie van beide.

Elk informatieproduct in de bedrijfszone is gebaseerd op gegevens uit de bronzone en/of de integratiezone. Soms zelfs op gegevens uit de bedrijfszone zelf (een ander informatieproduct, of het data-spoor).

N.B. Een informatieproduct is dus **niet per definitie** gebaseerd op een informatiegebied in de integratiezone. Als dat informatieproduct direct op de bronzone-data (eigenlijk: de bronzone-ontkoppelviews) kan worden gebaseerd heeft dat de voorkeur. [84]

In een informatieproduct worden gemaskeerde en ongemaskeerde gegevens nooit gecombineerd.

Er zijn daarom drie typen informatieproducten:

- Informatieproducten met gemaskeerde persoonsgegevens
  ("Gemaskeerde informatieproducten")
- Informatieproducten met ongemaskeerde persoonsgegevens
  ("Ongemaskeerde informatieproducten")
- Informatieproducten zonder persoonsgegevens

Soms bestaat er voor een informatieproduct een gemaskeerde en een ongemaskeerde variant; gelijk qua structuur en afleidingen, en met als enig verschil dat de eerste gebaseerd is op gemaskeerde gegevens, en de tweede op ongemaskeerde.

De **laadlogica informatieproducten** transformeert de gegevens uit de bronzone-ontkoppelviews (al dan niet gecombineerd met gegevens uit informatiegebieden in de integratiezone) en creëert op basis hiervan een informatieproduct.

De **laadlogica informatieproducten** wordt opgestart door de **master sequence informatievoorziening**. Dit is een overkoepelende, door een scheduler opgestarte, aansturingscomponent die zowel de **laadlogica integratiezone** (in de integratiezone) als het aanmaken van de erop gebaseerde informatieproducten (in deze zone) aanstuurt.

De **laadlogica informatieproducten** is, net als die van de **integratiezone**, maar beperkt te standaardiseren. Er wordt daarom waarschijnlijk geen gebruik gemaakt van harnassen; **maatwerk-ETL** is de standaard.
Deels wordt deze laadlogica overigens nog steeds gedreven worden door **stuur-metadata**.
Binnen het **maatwerk-ETL** wordt zoveel mogelijk gebruik gemaakt van standaard-bouwstenen, bijvoorbeeld t.b.v. logging. Ook generieke logica zoals voor het opbouwen/verversen van dimensies kan potentieel als bouwsteen worden ontwikkeld.

---

[83] In de PSA heet dit "gegevensleveringen". Omdat die term binnen DWH en Gegevensdiensten ook voor andere concepten wordt gebruikt, wordt in dit Conceptueel Ontwerp steeds de term "bestandsleveringen" gebruikt.

[84] E.e.a. conform het "kortste keten" principe

Sommige informatieproducten worden op basis van een abonnement (of vergelijkbare parameters) geleverd. Waar dit het geval is worden deze parameters pas tijdens het aanmaken van het informatieproduct (dus in de laadlogica informatieproducten) opgehaald, en dus niet langs de "reguliere" (via bronleveringen en bronzone) verwerkt en opgeslagen.[85]

Voor de meeste informatiegebieden blijft de uitvoer binnen het DIM; voor bestandsleveringen is dat echter niet het geval. De output daarvan komt terecht in een outbox-folder. Afhankelijk van het doel van de levering (een andere applicatie/serverlocatie of de KA-omgeving) is dat een server- of een transit-outbox.

De verwerkingsresultaten van de **laadlogica informatieproducten** (succes, ondervonden fouten, etc) komen in de **log-metadata**.

Voor bestandsleveringen en/of parameter-gedreven informatieproducten wordt daarnaast het **lever-spoor** bijgewerkt:

- Gebruikte parameters (en soms ook andere gebruikte gegevens) worden opgeslagen in het **data-spoor**
- Geleverde bestanden worden opgeslagen in het **archief bestandsleveringen**

De door de bronzone en integratiezone geleverde **actualiteits-metadata** is cruciaal bij het bepalen of aan alle randvoorwaarden voor het verversen van een informatieproduct is voldaan. Na het verversen van een informatieproduct wordt ook in de bedrijfszone de **actualiteits-metadata** bijgewerkt, zodat eenvoudig te achterhalen is tot op welk punt in de tijd een informatieproduct is bijgewerkt.

Het Data Lifecycle Management (DLM) van de gegevens in de diverse database gerelateerde informatieproducten (en in het data-spoor) wordt deels uitgevoerd door de **laadlogica informatieproducten**, en deels door de **DLM-tools bedrijfszone**.

*Ontwerp en implementatie van de **DLM-tools bedrijfszone** zijn onderdeel van fase 5. Meer details over deze tools zullen dus in een later stadium aan dit document worden toegevoegd.*

10.2 Gemaskeerde en ongemaskeerde informatieproducten

In een informatieproduct worden gemaskeerde en ongemaskeerde gegevens nooit gecombineerd.

Er zijn daarom drie typen informatieproducten te onderkennen:

- Informatieproducten met gemaskeerde persoonsgegevens ("Gemaskeerde informatieproducten")
- Informatiegebieden met ongemaskeerde persoonsgegevens ("Ongemaskeerde informatieproducten")
- Informatieproducten zonder persoonsgegevens

10.2.1 Gemaskeerde informatieproducten

Deze informatieproducten zijn gebaseerd op gemaskeerde bronzone-gegevens en/of gemaskeerde informatiegebieden, al dan niet gecombineerd met "overige gegevens" uit de bronzone en/of gegevens uit "informatiegebieden zonder persoonsgegevens".

N.B. Ongemaskeerde bronzone-gegevens en gegevens uit ongemaskeerde informatiegebieden mogen **niet** als input voor een gemaskeerd informatieproducten worden gebruikt.

[85] Zie paragraaf 4.2 (Gegevens als parameters).

10.2.2 Ongemaskeerde informatieproducten

Deze informatieproducten zijn gebaseerd op ongemaskeerde bronzone-gegevens en/of ongemaskeerde informatiegebieden, al dan niet gecombineerd met "overige gegevens" uit de bronzone en/of gegevens uit "informatiegebieden zonder persoonsgegevens".

N.B. Gemaskeerde bronzone-gegevens en gegevens uit gemaskeerde informatiegebieden mogen **niet** als input voor een ongemaskeerd informatieproduct worden gebruikt.

10.2.3 Informatieproducten zonder persoonsgegevens

Deze informatieproducten zijn óf uitsluitend gebaseerd op "overige gegevens" en/of gegevens uit "informatiegebieden zonder persoonsgegevens", óf bevatten aggregaten van persoonsgegevens (gemaskeerd of ongemaskeerd) die door die aggregatie anoniem zijn geworden (en dus niet meer als persoonsgegeven beschouwd hoeven te worden).

In dat tweede geval kunnen die aggregaten weer gecombineerd zijn met "overige gegevens" uit de bronzone en/of gegevens uit "informatiegebieden zonder persoonsgegevens".

N.B. Gemaskeerde en ongemaskeerde bronzone-gegevens, en gegevens uit gemaskeerde en ongemaskeerde informatiegebieden, mogen **alleen** als input voor een informatieproduct zonder persoonsgegevens worden gebruikt als ze bij het laden in dat informatieproduct door aggregatie anoniem worden gemaakt.

Het verdient de voorkeur om "anonieme" aggregaten van persoonsgegevens te baseren op gemaskeerde gegevens, zodat het DLM consistent blijft met andere voor rapportage/analyse bedoelde gegevens.

In twee gevallen kan hiervan afgeweken worden:

3. Het aggregaat is alleen relevant in combinatie met ongemaskeerde gegevens.
   In dat geval wordt het aggregaat, om consistent te blijven met de detailgegevens, in het algemeen ook gebaseerd op ongemaskeerde gegevens.
4. Het aggregaat wordt gebruikt in informatieproducten met twee varianten (zie volgende paragraaf).
   In dat geval wordt het aggregaat ook in twee varianten aangemaakt.

N.B. In alle gevallen (dus ook bij de standaardsituatie) dient in de precieze definitie van het aggregaat (zoals opgeslagen bij de "functionele termen" in de metadata) ook opgenomen te zijn op welke gegevens (gemaskeerd of ongemaskeerd) het aggregaat is gebaseerd, en dus welk DLM op het aggregaat van toepassing is.

10.2.4 Informatieproducten met twee varianten

Soms bestaan er van een informatieproducten een gemaskeerde en een ongemaskeerde variant; gelijk qua structuur en afleidingen, en met als enig verschil dat de eerste gebaseerd is op gemaskeerde data, en de tweede op ongemaskeerde.

Deze twee varianten worden ververst met dezelfde laadlogica; het enige verschil is de te gebruiken input-gegevens (gemaskeerd of ongemaskeerd) uit bronzone en/of integratiezone, en de te gebruiken locatie (in database of folder) van het product zelf. Deze keuze wordt middels stuur-metadata aan de laadlogica meegegeven.

N.B. Bij informatieproducten met twee varianten zullen eventuele bijbehorende aggregaten (zie vorige paragraaf) meestal ook in twee varianten moeten worden aangemaakt.

10.2.5 Traceerbaarheid via het lever-spoor

Het **lever-spoor** bevat de informatie die informatieproducten traceerbaar maakt, en bevat twee componenten:

- Het **data-spoor** (vergelijkbaar met de ANL-laag in DWH3) bevat voor elk parameter-gedreven informatieproduct de bij het aanmaken gebruikte parameters.
  Voor bestandsleveringen kunnen in het data-spoor ook de in het bestand geleverde gegevens (of de gegevens die gebruikt werden bij het aanmaken van het bestand) worden opgeslagen, bijvoorbeeld om ook "verschil-bestanden" te kunnen leveren. Of/hoe dat gebeurt is afhankelijk van de afnemerseisen, en zal dus per bestandslevering verschillen.
- Bij bestandsleveringen worden geleverde bestanden opgeslagen in het **archief bestandsleveringen**. Dit archief bestandsleveringen is overigens alleen bedoeld om herleveringen te kunnen ondersteunen; de afnemer (of de afnemende applicatie) is zelf verantwoordelijk voor de archivering conform Archiefwet.

10.3 Laadlogica informatieproducten (algemeen)

Deze paragraaf beschrijft alleen de aspecten die voor elk type informatieproduct relevant zijn. Zie de paragrafen 0 t/m 10.7 voor een meer gedetailleerde uitwerking per type informatieproduct.

10.3.1 Gedefinieerde gegevens, gedefinieerd eigenaarschap

Van elke gegevenselement in een informatieproduct (gekopieerd of afgeleid/gefilterd) moet duidelijk zijn wat de definitie ervan is, en wie (bron, afnemer, Gegevensdiensten) er de eigenaar van is.

Van alle gegevenselementen in de **informatieproducten** van de **bedrijfszone** dient duidelijk te zijn wat deze gegevenselementen betekenen en hoe deze gegevenselementen gevuld dienen te worden:

- Ieder gegevenselement in een **informatieproduct** moet gekoppeld zijn aan een functionele term afkomstig uit de functionele terminologie laag in IGC. Deze koppeling zal gemaakt worden in het datamodel van de bedrijfszone in IDA en raadpleegbaar zijn voor de eindgebruiker in IGC. Bij voorkeur dient de functionele term afkomstig te zijn uit een FUGEM, CGM of HUDSV. Als dit niet mogelijk is, bijvoorbeeld wanneer sprake is van een niet op HUDSV gebaseerd afgeleid veld, dan wordt een functionele term toegevoegd aan IGC.
  De definitie van een functionele term en de eigenaar van een functionele term worden vastgelegd in IGC.
- Voor ieder gegevenselement in een **informatieproduct** moet duidelijk gemaakt worden hoe deze kolom gevuld dient te worden. Dit zal veelal vanuit gegevenselementen vanuit de **bronzone** of **integratiezone** komen. De mapping wordt in IDA vastgelegd en zal, als onderdeel van het Functioneel Ontwerp, als basis dienen voor het ontwikkelwerk van de technische ontwikkelteams. Het resultaat zal voor de eindgebruikers raadpleegbaar zijn in IGC.

Ook van het informatieproduct zelf moet duidelijk zijn wie de eigenaar is, en dus verantwoordelijk voor specificatie en acceptatie van product en bijbehorende laadlogica.

10.3.2 Maatwerk, maar wel met een standaard-aanpak

Hoewel de detailstappen voor elk informatieproduct verschillen zijn er ook een aantal stappen en stukken functionaliteit die voor alle informatieproducten (of tenminste voor alle informatieproducten van hetzelfde type) gelijk zijn.

De altijd te doorlopen stappen zijn (voor alle fysieke informatieproducten, dus niet voor gegevensvensters):

1 Start
Haal, m.b.v. de door de master sequence meegegeven parameters, de voor het informatieproduct benodigde stuur-metadata op.

Maak een record aan in de log-metadata met als status "lopend".

2 Initialisatie
Bepaal, op basis van DataStage omgevingsvariabelen, de omgevingsafhankelijke[86] delen van paden en bestandsfolders.

3 Detecteer beschikbaarheid input-gegevens
Op basis van de actualiteits-metadata.

4 Haal, voor parameter-gedreven informatieproducten, de parameters op, en creëer of ververs het informatieproduct (en, voor bestandsleveringen en/of parameter-gedreven leveringen, het **data-spoor**). Sla bestandsleveringen ook op in het **archief bestandsleveringen.**

X1 Verzend notificaties
Meld succes of issues via e-mail aan de relevante partijen.
Welke meldingen naar welke partijen moeten, en welke emailadressen voor deze partijen gebruikt moeten worden, staat in de stuur-metadata.

X2 Einde job
Vervang, in de log-metadata, de markering "lopend" door "klaar" of "afgebroken", afhankelijk van het resultaat van de verwerking tot op dit punt.
Werk de actualiteits-metadata bij.
[einde ETL-job]

Binnen het **maatwerk-ETL** wordt zoveel mogelijk gebruik gemaakt van steeds hetzelfde stappenplan en steeds dezelfde standaard-bouwstenen (bijvoorbeeld t.b.v. logging).
Dit stappenplan wordt echter waarschijnlijk niet standaard afgedwongen door een harnas of ander ETL-raamwerk.

10.3.3 Tijd-gedreven of gegevens-gedreven verversing

Afhankelijk van de vereisten van de afnemers moeten informatieproducten tijd-gedreven of gegevens-gedreven ververst worden:

- **Tijd-gedreven**
  Het informatieproduct moet op een vast moment beschikbaar zijn, gebaseerd op de op dat moment beschikbare input-gegevens.
  Zijn die gegevens (deels) nog niet ververst (of loopt die verversing nog), dan wordt het informatieproduct gebaseerd op de meest recente bruikbare set. Afhankelijk van de consistentie-eisen aan het informatieproduct kan dit betekenen dat voor één bron minder actuele gegevens worden gebruikt, of voor alle bronnen.

[86] Met "omgevingsafhankelijk" wordt bedoeld: afhankelijk van de OTAP-omgeving waarin het proces draait.

- **Gegevens-gedreven**
  Het informatieproduct moet ververst worden zodra de input-gegevens ervoor dat ook zijn.

Voor tijd-gedreven informatieproducten haalt de laadlogica de input-gegevens, waar nodig, op met een filter op de administratieve tijdlijn.

Voor gegevens-gedreven informatieproducten controleert de laadlogica in de stap "Detecteer beschikbaarheid input-gegevens" in de actualiteits-metadata of alle vereiste gegevens ververst zijn, en blijft, indien nodig, deze metadata "pollen" tot zulks het geval is.

10.3.4 Laden van informatieproducten met twee varianten

Onderstaande is alleen relevant voor fysieke informatieproducten, dus niet voor gegevensvensters.

De ETL-job voor het vullen van een gemaskeerde of ongemaskerde variant van een informatieproduct zal dezelfde zijn. Dit kan doordat de input-gegevens en de output-gegevens qua gegevensstructuur en tabel/view/kolom-namen identiek zijn voor zowel het gemaskeerde als het ongemaskeerde informatieproduct. Het verschil zit in de schema-namen. Hierdoor is het mogelijk deze als parameters voor de ETL-job te gebruiken.

10.3.5 Verbergen zeer gevoelige gegevens

Afhankelijk van de rechtsgrond/proportionaliteit/subsidiariteit van het achterliggend proces moeten zeer gevoelige personen (VIP's), zeer gevoelige zaken (BZ, BGB) en zeer gevoelige waarden (bv. de reden van een detentie) vaak uit een informatieproduct worden verwijderd.[87]

Bij informatieproducten met twee varianten (zie vorige paragraaf) zal dit, in het algemeen, alleen gelden voor de ongemaskeerde variant.

In een aantal gevallen dienen sommige afnemers van het informatieproduct wél toegang te hebben tot gevoelige personen en/of zaken, en andere niet.
Waar mogelijk wordt dit geregeld door twee gegevensvensters op het informatieproduct te leggen; één met, en één zonder zeer gevoelige personen/zaken. Is dit niet mogelijk, dan moet de filtering binnen de end-user zone worden afdwongen (bv. in een BusinessObjects-universe).

In beide gevallen dient het informatieproduct zo te zijn ingericht dat het uitfilteren van zeer gevoelige gegevens zo eenvoudig mogelijk is.

Voor bestandsleveringen is het ook mogelijk om de leveringen in twee smaken (en op twee locaties) aan de afnemers te leveren: met en zonder zeer gevoelige personen/zaken.[88]

10.3.6 Parameter-gedreven informatieproducten

In theorie zijn er twee typen parameters relevant bij het aanmaken van informatieproducten[89]:

- Abonnementsgegevens
- Ad hoc parameters

Op dit moment zijn er nog geen informatieproducten o.b.v. ad hoc parameters onderkend, en slechts één o.b.v. abonnementsgegevens: de SUAG-levering.

---

[87] Zie paragraaf 4.1.3.1~~4.1.3~~ (Zeer gevoelige personen/zaken) en 4.1.3.2~~4.1.4~~ (Velden met zeer gevoelige waarden)

[88] Meer informatie over VIP-, EP- en BZ-filtering in Bijlage J: Filtering op VIP's en BZ/EP+

[89] Zie paragraaf 4.2 (Gegevens als parameters).

Abonnementsgegevens staan, voor het Polisdomein, standaard in UGC. Daar haalt ook de SUAG-levering zijn parameters grotendeels[90] vandaan.

Er zijn plannen (binnen de divisie Gegevensdiensten, maar buiten het project en/of de afdeling DWH) om UGC breder in zetten, en er ook de contracten van leveringen aan en leveringen uit het DIM in op te nemen.

Tot nader order zal de SUAG-levering uit het DIM zijn parameters op dezelfde wijze ophalen als de bestaande SUAG-levering uit DWH3 (via, door UGC en Polis+ geleverde, web services).

Als er meer bekend is over die bredere inzet van UGC (naar verwachting eind fase 5 of later), dan zal hiervoor een standaard-interactie met het DIM worden gedefinieerd.

~~Tot~~[GL(14]~~ nader order zal de SUAG-levering uit het DIM zijn parameters op dezelfde wijze ophalen als de bestaande SUAG-levering uit DWH3 (via, door UGC en Polis+ geleverde, web services).~~

10.3.7 Informatieproducten met een "bron van verbeteringen"

Voor een beperkt aantal informatieproducten wordt, door de afnemer en buiten het DIM, een gegevenskwaliteitsanalyse uitgevoerd voordat het product breder verspreid wordt.
Een voorbeeld hiervan is de SUAG-levering.

Bij kwaliteitsproblemen moeten gegevens die niet in de bron kunnen worden aangepast via een andere weg worden verbeterd.
Om te garanderen dat hierbij de traceerbaarheid in stand blijft gebeurt dit m.b.v. een "bron van verbeteringen".[91]

In principe doorloopt het DIM voor een "bron van verbeteringen" de standaard-verwerkingsprocessen in ontsluitingszone en bronzone. Wel kan het nodig zijn om deze processen "ad hoc" aan te kunnen roepen (dus niet alleen via de "hartslag"-scheduling van het DIM), zodat het reparatieproces zoveel mogelijk interactief blijft.

10.4 Datamarts

Een datamart is een op maat gemaakte gegevensverzameling, vaak (maar niet altijd!) dimensioneel gestructureerd, ten behoeve van specifieke gebruikersgroepen en/of -processen (meestal rapportages, OLAP, dashboards, of vergelijkbaar), inclusief self service BI.

Vaak vereist dit het op maat maken van complexe transformaties en afleidingen.

10.4.1 Scoping van een datamart

10.4.1.1 Logisch samenhangend, soms deels virtueel

Net als een informatiegebied bevat een datamart een logisch samenhangende groep afgeleide en/of geïntegreerde gegevens, gebaseerd op één of meer bronnen.

Eventuele voor de datamart benodigde conforme dimensies kunnen "virtueel" in de datamart worden opgenomen door een (meestal gematerialiseerde) view op die dimensie (zoals beschikbaar in de integratiezone) op te nemen in het schema van de datamart.
Belangrijk is dat die view dan zo is ingericht dat de grant op de onderliggende tabel in de

[90] En deel van de parameters komt uit Polis+. Dat is oneigenlijk gebruik van die applicatie, maar valt buiten het aandachtsgebied van het DIM, de afdeling DWH en/of het project DataFabriek.

[91] Zie Bijlage G: ~~Bijlage F:~~ Datakwaliteitsbeheer, onder "Overschrijven kan niet, verbeteren wel"

integratiezone wordt "geërfd" van de grant op de view, zodat het toegangsbeheer voor de datamart eenvoudig blijft.

N.B. Als de gegevens in het DIM zonder verdere voorbewerking (dus zoals beschikbaar in de ontkoppelviews op de bronzone en/of de informatiegebieden in de integratiezone) bruikbaar zouden zijn voor de afnemer, dan is een datamart **niet** noodzakelijk; conform het "kortste keten" principe is het dan beter om de afnemer toegang te geven middels een gegevensvenster.
Alleen als dit tot onoverkomelijke bezwaren in de historieafhandeling leidt wordt hiervan afgeweken.

10.4.1.2 Gegevenstechnisch samenhangend

Als de logische samenhang (zie vorige paragraaf) leidt tot een datamart met, binnen die datamart, zeer verschillende gegevensstructuren, historisch gedrag en/of detail-niveaus, dan is het vaak handig om die datamart in meerdere datamarts op te knippen.

Opknippen is ook handig als de bewaartermijnen voor de gegevens binnen de datamart sterk verschillen.

10.4.1.3 Niet per definitie in volledig detail

Anders dan een informatiegebied (dat meerdere informatieproducten moet kunnen ondersteunen) worden gegevens niet per definitie in volledig detail in een datamart opgenomen.

Het in de datamart op te nemen detail volgt uit de afnemers-vereisten, die op hun beurt weer afhankelijk zijn van de, bij/door de afnemer uitgevoerde, controle op rechtsgrond, proportionaliteit en subsidiariteit.

10.4.1.4 Gemaskeerd en ongemaskeerd gescheiden

Een datamart mag nooit zowel gemaskeerde als ongemaskeerde persoonsgegevens bevatten. Lijkt hiervoor toch een noodzaak te bestaan, dan is dat waarschijnlijk te wijten aan non-compliant afnemerseisen en/of een verkeerd ontworpen datamart.

Soms bestaat er voor een datamart wél een gemaskeerde en een ongemaskeerde variant; gelijk qua structuur en afleidingen, en met als enig verschil dat de eerste gebaseerd is op gemaskeerde gegevens, en de tweede op ongemaskeerde.

Bij toegang tot de datamart via BusinessObjects wordt hierbij, in het ideale geval één universe gebruikt, dat, afhankelijk van de gebruikersrechten, de gemaskeerde of de ongemaskeerde variant van de datamart benadert.

Gezamenlijk met de lijn wordt in fase 4 uitgezocht of dit mogelijk [GL(15)]*is. Is dit niet het geval, dan zijn twee universes noodzakelijk.*

10.4.1.5 Met en zonder zeer gevoelige gegevens

Afhankelijk van de rechtsgrond/proportionaliteit/subsidiariteit van het achterliggend proces mogen zeer gevoelige personen, zaken en waarden vaak niet in een datamart worden getoond.

Bij datamarts met twee varianten zal dit, zoals bij elk informatieproduct, in het algemeen alleen gelden voor de ongemaskeerde variant.

In een aantal gevallen dienen sommige afnemers van de datamart wél toegang te hebben tot zeer gevoelige personen en/of zaken, en andere niet.

Waar mogelijk wordt dit geregeld door twee gegevensvensters op de datamart te leggen; één met, en één zonder zeer gevoelige personen/zaken, en de BI-tooling niet direct op de datamart aan te sluiten, maar op de gegevensvensters.

Bij toegang tot de datamart via BusinessObjects wordt hierbij in het ideale geval, vergelijkbaar met de toegang tot gemaskeerde of ongemaskeerde gegevens, één universe gebruikt, dat, afhankelijk van de gebruikersrechten, het DIM benaderd via het gegevensvenster mét of het gegevens zonder zeer gevoelige gegevens.

Gezamenlijk met de lijn wordt in fase 4 uitgezocht of dit mogelijk [GL(16]*is. Is dit niet het geval, dan moet de filtering binnen het universe worden afdwongen.*

N.B. Het inbouwen van het VIP-filter in het laadproces van de datamart (of het tijdens dat laadproces creëren van een "VIP J/N"-attribuut t.b.v. filtering in gegevensvenster of universe) is alleen toegestaan als de persoonsdimensie (of equivalent) van die datamart bij iedere verversing opnieuw wordt opgebouwd. Dit omdat VIP-filters "met terugwerkende kracht", en dus ook op de historische gegevens en/of afgesloten administratie versies moeten werken.[92] De "hartslag" van de datamart moet daarnaast relatief hoog zijn (dag of week).

10.4.2 Gegevensafhandeling binnen een datamart

10.4.2.1 DLM datamart t.o.v. de bronzone en de integratiezone

De bronzone dient als stabiele databasis. Hierdoor kan, binnen de bewaartermijnen van die bronzone, een datamart altijd weer opnieuw opgebouwd worden. Dit geldt ook voor de eventueel door die datamart gebruikte informatiegebieden uit de integratiezone.

In principe zal een datamart bij verversen op een slimme manier opnieuw worden opgebouwd tenzij er een functionele vereiste is om historische rapportageresultaten constant te houden, en deze vereist niet kan worden gedekt met een versionering o.b.v. de administratieve tijdlijn.

Zie paragraaf 14.5 (DLM en overige archivering/vernietiging in de bedrijfszone) voor een gedetailleerde beschrijving van het DLM binnen de bedrijfszone.

10.4.2.2 Gebruikte modellerings-concepten

De modellering van een datamart is afhankelijk van de afnemers-vereisten.

Dimensionele modellering zal het meeste voorkomen, maar ook 3NF en gedenormaliseerde "platte" structuren kunnen voorkomen.

10.4.2.3 Versionering

Ook de versionering binnen een datamart is afhankelijk van de afnemers-vereisten. Er is dus ook geen standaard deltaproces.

Voor gemaskeerde datamarts zal dit meestal versionering langs twee tijdlijnen (administratief en geldigheid) zijn. Dit om reproduceerbaarheid van S&V-rapportages te kunnen ondersteunen.

[92] Meer informatie over VIP-, EP- en BZ-filtering in Bijlage J: Filtering op VIP's en BZ/EP+

Versionering enkel op geldigheid (dus gebaseerd op alleen de administratief actieve gegevens uit de bronzone), of zelfs geen enkele versionering (alleen administratief én qua geldigheid actieve gegevens) is mogelijk, met name bij ongemaskeerde datamarts.

Voor beide tijdlijnen geldt dat de versionering versimpeld kan zijn tot alleen "snapshots" op de rapportage-momenten (b.v. maand ultimo).

N.B. In alle gevallen dient in de precieze definities van (de gegevenselementen in) de datamart (zoals opgeslagen bij de "functionele termen" in de metadata) ook opgenomen te zijn hoe de gegevens geversioneerd zijn.

10.4.2.4 Parameter-gedreven datamarts

Naar verwachting zullen parameter-gedreven datamarts niet voorkomen.
Mocht dat toch het geval zijn, dan zijn deze datamarts waarschijnlijk een tussenstap bij het aanmaken van een bestandslevering.[93]

10.4.2.5 Beschikbaarheid tijdens verversen

Afhankelijk van de beschikbaarheidseisen van de afnemer (zoals vastgelegd in de SNO[94]) en de omvang van de datamart kan een datamart:

- Tijdens het verversen niet beschikbaar zijn
- Tijdens het verversen wel beschikbaar zijn, maar zonder gegarandeerde consistentie. De SNO bevat dan waarschijnlijk een tijdsvenster waarbinnen die consistentie wél gegarandeerd wordt (en er dus niet ververst mag worden)
- Omhuld worden door een gematerialiseerd gegevensvenster, zodat tijdens het verversen de oude situatie (van net voor verversen) voor de afnemers beschikbaar blijft

10.5 Bestandsleveringen

Een bestandslevering is een levering van in meer of mindere mate bewerkte gegevens, doorgaans in de vorm van platte bestanden. Voorbeelden zijn leveringen aan pensioenfondsen[95] en aan analyse-omgevingen.

Bestandsleveringen kunnen aan andere applicaties (of server-locaties) zijn, of aan personen/afdelingen.

10.5.1 Scoping van een bestandslevering

10.5.1.1 Gemaskeerd en ongemaskeerd gescheiden

Een bestandslevering mag nooit zowel gemaskeerde als ongemaskeerde persoonsgegevens bevatten. Lijkt hiervoor toch een noodzaak te bestaan, dan is dat waarschijnlijk te wijten aan non-compliant afnemerseisen.

Soms bestaat er voor een bestandslevering wél een gemaskeerde en een ongemaskeerde variant; gelijk qua structuur en afleidingen, en met als enig verschil dat de eerste gebaseerd is op gemaskeerde gegevens, en de tweede op ongemaskeerde.

---

[93] Paragraaf 10.5 (Bestandsleveringen).

[94] ServiceNiveauOvereenkomst; Nederlands voor Service Level Agreement

[95] N.B. Alle leveringen van het DIM naar partijen **buiten** UWV lopen vanaf het DIM ~~is~~ eerst naar een partij **binnen** UWV. Vanuit het DIM gezien zijn alle leveringen dus "intern"

Dit zal met name kunnen voorkomen bij bestandsleveringen aan analyseomgevingen buiten het DIM (bijvoorbeeld DMAP).

De wijze waarop de levering aan DMAP moet worden ingericht is afhankelijk van, buiten [GL(17)]*het project op te stellen, standaarden voor analyse-omgevingen. Als die standaarden er in fase 5 nog niet zijn, dan zal levering aan DMAP de huidige (ongemaskeerde) praktijk voortzetten.*

10.5.1.2 Parameter-gedreven bestandsleveringen

Parameter-gedreven informatieproducten zullen meestal bestandsleveringen zijn.

Op dit moment is er één parameter-gedreven bestandslevering in scope: de SUAG-levering. Zie paragraaf 10.3.6 (Parameter-gedreven informatieproducten~~Parameter-gedreven informatieproducten~~) voor de aannames voor dit informatieproduct.

10.5.1.3 Bestandsleveringen o.b.v. datamarts

De input-gegevens voor een bestandslevering zullen meestal afkomstig zijn uit bronzone en integratiezone. Bestandsleveringen o.b.v. een datamart zijn echter ook mogelijk.

Dit is met name handig bij bestandsleveringen waarvoor hoge traceerbaarheids- en/of datakwaliteitseisen gelden; alle input-gegevens voor de levering staan dan eenvoudig analyseerbaar in de datamart.

N.B. Als een dergelijk bestandslevering parameter-gedreven is, dan is de onderliggende datamart dat waarschijnlijk ook.

10.5.1.4 Met of zonder zeer gevoelige gegevens

Het al dan niet moeten verwijderen van zeer gevoelige personen, zaken en waarden uit een bestandslevering volgt uit de afnemers-vereisten, die op hun beurt weer afhankelijk zijn van de, bij/door de afnemer uitgevoerde, controle op rechtsgrond, proportionaliteit en subsidiariteit.

10.5.2 Gegevensafhandeling binnen een bestandslevering

10.5.2.1 Gebruik van de outbox

Bestandsleveringen kunnen aan andere applicaties (of server-locaties) zijn, of aan personen/afdelingen.

In het eerste geval komt het aangemaakte bestand in een **server-outbox**, waarna het kan worden verzonden naar (of opgehaald door) de afnemende applicatie.
Verzenden/ophalen gebeurt conform de Gegevensdiensten-standaarden hiervoor. Deze zijn ten tijde van het schrijven van dit ontwerp nog niet volledig uitgekristalliseerd.

In het tweede geval komt het aangemaakte bestand in een **transit-outbox** (fysiek geplaatst op de server, maar, bijvoorbeeld via UCRA, toegankelijk vanuit de KA-omgeving), waarvandaan het kan worden opgehaald door KA-users.

Elke bestandslevering (of elke groep bestandsleveringen met identieke afnemers) heeft zijn eigen outbox.

10.5.2.2 Gebruik van het lever-spoor

Bij bestandsleveringen wordt het geleverde bestand opgeslagen in het **archief bestandsleveringen**.

Indien nodig worden ook de gegevens in het bestand (of de gegevens die punt waren bij het aanmaken van het bestand) in gestructureerde vorm in **het data-spoor** opgeslagen.

Ook eventuele invoer-parameters worden, t.b.v. traceerbaarheid en reproduceerbaarheid, in dit data-spoor opgeslagen.

10.5.2.3 Archivering en vernietiging van bestandsleveringen

Het **archief bestandsleveringen** is alleen bedoeld om herleveringen te kunnen ondersteunen; de afnemer (of de afnemende applicatie) is zelf verantwoordelijk voor de archivering conform Archiefwet.

De bewaartermijn in dit archief is vastgelegd in de SNO en zal, naar verwachting ongeveer gelijk zijn aan de maximale periode tussen twee bestandsleveringen.

Zie paragraaf 14.5 (DLM en overige archivering/vernietiging in de bedrijfszone) voor een gedetailleerde beschrijving van het DLM binnen de bedrijfszone.

10.6 Gegevensvensters

Een gegevensvenster is een virtuele gegevensverzameling (deelverzameling van de in het DIM beschikbare gegevens) die ontsloten wordt ten behoeve van self-service BI en/of self service analytics, d.w.z. rechtstreeks gebruik door afnemers, bijvoorbeeld t.b.v. querying en ad-hoc rapportage.
Anders dan bij datamarts is er geen sprake van complexe transformaties en afleidingen. Wel bevatten gegevensvensters filters op tabel-, rij- en attribuutniveau. Dit om te garanderen dat de getoonde deelverzameling alleen gegevens bevat waarvoor rechtsgrond geldt, en waarvan het gebruik voldoet aan de AVG-principes m.b.t. proportionaliteit en subsidiariteit..

Technisch gezien is een gegevensvenster niet meer dan een database-schema, met daarin een verzameling "domme" Oracle-views.

Belangrijk is overigens dat die views zo zijn ingericht dat de grant op de onderliggende tabellen "geërfd" wordt van de grant op de gegevensvenster-view (eigenlijk: het gegevensvensterschema). Hierdoor blijft het toegangsbeheer voor gegevensvensters zo eenvoudig en veilig mogelijk:

- Autorisatie van een afnemer op het gegevensvenster zelf voldoet; het gegevensvenster zelf is weer geautoriseerd op de onderliggende tabellen.
- De afnemer heeft alleen toegang tot het gegevensvenster, en géén toegang tot de onderliggende tabellen.

10.6.1 Scoping van een gegevensvenster

10.6.1.1 Logisch óf qua doelproces samenhangend

Gegevensvensters zijn vooral bedoeld om de controles op rechtsgrond, proportionaliteit en subsidiariteit (inclusief de daaruit volgende autorisatieprocessen) te vereenvoudigen. Daarom omvat een gegevensvenster niet altijd uitsluitend logisch samenhangende gegevens; als een doelproces meerdere groepen logisch samenhangende gegevens vereist, dan is het handiger om al die groepen in één gegevensvenster samen te voegen.

Dit vereenvoudigt niet alleen het autorisatieproces, maar maakt het ook eenvoudiger om, bij controles op rechtsgrond/proportionaliteit/subsidiariteit, te kijken naar de gevoeligheid van de totale gebruikte dataset.

10.6.1.2 Gefilterd op vereiste

Een gegevensvenster toont alleen de tabellen, rijen en attributen die voor het doelproces noodzakelijk zijn. E.e.a. conform de afnemers-vereisten, die op hun beurt weer afhankelijk zijn van de, bij/door de afnemer uitgevoerde, controle op rechtsgrond, proportionaliteit en subsidiariteit.

Dat betekent ook dat een gegevensvenster-view altijd een selectie op specifieke kolommen moet bevatten, en geen "SELECT *".

10.6.1.3 Toegang tot alle zones

Gegevensvensters kunnen toegang bieden tot gegevens uit de bronzone en de integratiezone, en tot andere informatieproducten (of het data-spoor) uit de bedrijfszone.

Omgekeerd geldt ook: directe ("virtuele") toegang tot de bronzone-ontkoppelviews en/of de informatiegebieden in de informatiezone is alleen mogelijk via gegevensvensters.

Daardoor blijft de bedrijfszone de enige zone die afnemers toegang tot de gegevens in het DIM biedt.

10.6.1.4 Overlap is regel, geen uitzondering

De scoping van een gegevensvenster is gericht op het afnemende proces. Als meerdere afnemende processen dezelfde gegevens vereisen, dan zullen gegevensvensters elkaar dus overlappen.

10.6.1.5 Geen "virtuele datamart"

De views binnen een gegevensvenster bevatten alleen simpele filters, en geen transformaties en/of joins (m.u.v. een eventuele join met de VIP-lijst om deze uit te filteren).[96]

Als de afnemer wél transformaties vereist, dan dienen die in een informatiegebied en/of een datamart te worden opgenomen. Het gegevensvenster kan dan weer wel toegang bieden tot dat informatiegebied of die datamart.

N.B. De PSA beschrijft ook "virtuele datamarts". Die zijn echter meer een toekomstvisioen, gebaseerd op, nog niet binnen UWV beschikbare, tools voor data virtualisatie.

10.6.1.6 Gemaskeerd en ongemaskeerd gescheiden

Een gegevensvenster mag nooit zowel gemaskeerde als ongemaskeerde persoonsgegevens bevatten. Lijkt hiervoor toch een noodzaak te bestaan, dan is dat waarschijnlijk te wijten aan non-compliant afnemerseisen.

Soms bestaat er voor een gegevensvenster wél een gemaskeerde en een ongemaskeerde variant; gelijk qua structuur en afleidingen, en met als enig verschil dat de eerste gebaseerd is op gemaskeerde gegevens, en de tweede op ongemaskeerde.

Dit zal met name kunnen voorkomen bij gegevensvensters t.b.v. analyseomgevingen buiten het DIM (bijvoorbeeld DMAP).

De wijze waarop de levering aan DMAP moet worden ingericht is afhankelijk van, buiten het project op te stellen, standaarden [GL(18]*voor analyse-omgevingen. Als die standaarden er in fase 5 nog niet zijn, dan zal levering aan DMAP de huidige (ongemaskeerde) praktijk voortzetten.*

[96] Meer informatie over VIP-, EP- en BZ-filtering in Bijlage J: Filtering op VIP's en BZ/EP+

10.6.1.7 Met en zonder zeer gevoelige gegevens

Afhankelijk van de rechtsgrond/proportionaliteit/subsidiariteit van het achterliggend proces moeten zeer gevoelige personen, zaken en waarden vaak uit een datamart worden verwijderd.

Bij gegevensvensters met twee varianten zal dit, zoals bij elk informatieproduct, in het algemeen alleen gelden voor de ongemaskeerde variant.

In een aantal gevallen dienen sommige afnemers van een gegevensvensters wél toegang te hebben tot zeer gevoelige personen en/of zaken, en andere niet.

Dit wordt geïmplementeerd als een "gegevensvenster op een gegevensvenster", waarbij het tweede gegevensvenster de zeer gevoelige gegevens uit het eerste filtert.

10.6.2 Gegevensafhandeling binnen een gegevensvenster

10.6.2.1 Materialiseren of niet

De enige gegevensafhandeling binnen een gegevensvenster is het al dan niet materialiseren van (een deel van) de erin opgenomen views.

N.B. Gegevensvenster-views op bronzone-gegevens hoeven nooit gematerialiseerd te worden, aangezien de bronzone-ontkoppelviews waarop het venster dan gebaseerd is zelf al gematerialiseerd zijn.

10.7 Zandbakken

Een zandbak is een data-omgeving t.b.v. analyse.
Een zandbak kan virtueel of fysiek zijn, of een combinatie van beide.

Een zandbak binnen het DIM wordt (op dit moment) geïmplementeerd als een gegevensvenster, met daarin vaak ook informatiegebieden en/of (al dan niet in 3NF-gemodelleerde) datamarts.

Een zandbak buiten het DIM wordt door het DIM voorzien van gegevens middels een gegevensvenster of een verzameling bestandsleveringen.

De regels/oplossingen voor zandbakken zijn daarmee direct af te leiden uit die voor gegevensvensters, datamarts en bestandsleveringen.

N.B. In de toekomst, als UWV ook de beschikking heeft over technologie voor niet-relationele zandbakken, zal een zandbak wél een apart geïmplementeerd informatieproduct worden. Dit is ook het geval als Gegevensdiensten fysieke zandbakken (ook relationele) gaat hosten. Voor beide gevallen is nog geen planning of detail beschikbaar.

10.8 DLM-tools bedrijfszone

Het Data Lifecycle Management van de informatieproducten in de bedrijfszone wordt deels uitgevoerd door de **laadlogica informatieproducten**, en deels door de **DLM-tools bedrijfszone**.

*Deze paragraaf is een placeholder; ontwerp en implementatie van de **DLM-tools bedrijfszone** zijn onderdeel van fase 5. Meer details over deze tools zullen dus in een later stadium aan dit document worden toegevoegd.*

Zie paragraaf 14.5 (DLM en overige archivering/vernietiging in de bedrijfszone) voor meer informatie.

10.9 Technische aspecten

10.9.1 Database-inrichting

Elk informatieproduct heeft een eigen databaseschema om het toegangsbeheer en de beveiliging te vergemakkelijken.

Autorisatie op informatieproducten gebeurt logisch op schema-niveau; een afnemer/rol heeft óf toegang tot alle tabellen/views in het schema, óf tot geen enkele.

Elke schemanaam heeft een suffix die aangeeft of het een informatieproduct met gemaskeerde persoonsgegevens, met ongemaskeerde persoonsgegevens, of met overige gegevens (geen persoonsgegevens) betreft.
Voor een datamart en/of een gegevensvenster met twee varianten zijn alle tabel- en kolomnamen (resp. view- en kolomnamen) voor deze twee varianten gelijk, en verschilt in de schemanaam alleen de suffix.

Alle binnen de bedrijfszone (in datamarts en gegevensvensters) gebruikte views zijn zo ingericht dat de grant op de onderliggende tabellen "geërfd" wordt van de grant op de bedrijfszone-view (eigenlijk: het gegevensvensterschema). Dit om te voorkomen dat voor autorisatie op een informatieproduct kennis van de structuur en/of input-gegevens van dat informatieproduct vereist is.

10.9.2 Toegangsrechten

Alle folders en database-objecten in de bedrijfszone zijn (op A en P) alleen toegankelijk voor geautomatiseerde processen en (read-only) voor geautoriseerde eindgebruikers.

Hierop zijn twee uitzonderingen:

- DBA's kunnen folders aanmaken en database-objecten creëren/wijzigen t.b.v. het creëren van nieuwe of het wijzigen van bestaande informatieproducten.
  Idealiter loopt dit overigens via een door een release-tool uitgevoerd script, zodat deze toegang niet noodzakelijk is
- DBA/Beheerders/ontwikkelaars kunnen, in geval van een calamiteit, tijdelijk toegang krijgen tot database-objecten t.b.v. onderzoek. Deze tijdelijke toegang verloopt via een "red envelope"-procedure.[97]

Deze toegangsrechten zijn "set-gebaseerd"; een eindgebruiker krijgt toegang tot een volledig informatieproduct. Toegang kan hierbij verleend worden door de eindgebruiker zélf te autoriseren op een informatieproduct, of door de eindgebruiker in een BI-tool rechten te geven opeen "connectie" (waarvan het wachtwoord niet zichtbaar is), en die connectie de rechten te autoriseren op het informatieproduct.[98]

Een bijzonder geval zijn informatieproducten die in een gefilterde en een ongefilterde variant voorkomen.

Een voorbeeld:

> Een datamart bevat uitkeringsgerechtigden met hun VIP-status.
> Een gegevensvenster op die datamart bevat een filter op die VIP's. Er zijn nu twee datasets:
>
> - Eén met alle uitkeringsgerechtigden met hun VIP-status

[97] Zie hoofdstuk 18 (Informatiebeveiliging en –beheer) voor meer informatie over deze procedure.

[98] Voor BusinessObjects is dit uitgewerkt in de BO-standaarden (project-deliverable P187)

- Eén met alle niet-VIP uitkeringsgerechtigden met hun VIP-status (die dus altijd de waarde "geen VIP" heeft)[99]

Een eindgebruiker wordt vervolgens geautoriseerd op óf de dataset ~~,~~mét VIP's of die zonder (of op geen van beide, natuurlijk). In beide gevallen is dat een volledig informatieproduct.~~.~~

10.9.3 Technische velden

Niet alle tabellen in de datamarts bevatten dezelfde technische velden. Dit omdat de modellering/versionering per informatiegebied kan verschillen.

Technische velden met dezelfde functie zullen echter wel altijd, in alle informatieproducten, dezelfde naam en afleiding hebben.

Deze namen en afleidingen zijn dezelfde als die voor informatiegebieden (in de integratiezone), en zijn daarvoor beschreven in paragraaf 9.6.3 (Technische velden~~Technische velden~~).

De technische namen worden in de technische documentatie vastgelegd, op basis van de standaarden en richtlijnen op dit gebied.

10.10 Impact ontwerpuitgangspunten

De ontwerpuitgangspunten zijn meegenomen in (dit deel van) het conceptueel ontwerp. Daarnaast moeten ze meegenomen worden in de detaillering en de realisatie van dat ontwerp.

Ontwerpuitgangspunt	**Impact**
Bewezen concepten	Alle typen informatieproducten zijn "market best practice", met uitzondering van de, technisch niet risicovolle, gegevensvensters.
Geen realtime ambitie	Alle verwerking is batch gedreven, Geen informatieproducten met "near real time" / "24x7" service levels.
Maximale ontkoppeling	Het verversen van informatieproducten is niet hard gekoppeld aan de levering van de vereiste bronnen en/of informatiegebieden, maar gescheduled op DIM-hartslag, met vervolgens een eigen controle op de beschikbaarheid van de benodigde input-gegevens.
Specifiek waar mogelijk, gemeenschappelijk waar noodzakelijk of waardevol	Generieke bouwstenen worden hergebruikt Sturing middels standaard master sequences.
Kortste keten	Geen informatiegebieden als input gebruiken als het informatieproduct ook met bronzone-data toe kan. Geen datamart bouwen als een gegevensvenster voldoet.

[99] Meer informatie over VIP-, EP- en BZ-filtering in Bijlage J: Filtering op VIP's en BZ/EP+

Ontwerpuitgangspunt	**Impact**
Compliant	Strikte scheiding van gemaskeerde en ongemaskeerde gegevens; een informatieproduct mag maar één van beide bevatten. Toegang van eindgebruikers tot de DIM-data altijd via een informatieproduct in de bedrijfszone. Controleprocessen m.b.t. rechtsgrond, proportionaliteit en subsidiariteit (AVG) daardoor eenvoudig in te richten. DLM bedrijfszone conform AVG én Archiefwet. Voldoende rijk (qua attributen) om rij-filters (in informatieproduct) eenvoudig te kunnen implementeren.
Eenvoudig	Er worden in de verwerking geen extra stappen gedaan indien er geen functionele noodzaak voor is. De functionaliteit is opgedeeld in logische stappen, met optimaal gebruik van herbruikbare bouwstenen.
Gebruiksvriendelijk	Alleen informatieproducten voor specifieke afnemers-groepen of processen; geen generieke informatieproducten. Datamodellering niet gestandaardiseerd, maar geoptimaliseerd voor een specifiek informatieproduct.
Gedefinieerd	Eigenaarschap definities/afleidingen belegd (bij bron, Gegevensdiensten of afnemer) en vastgelegd. Afleidingen gevalideerd/geaccordeerd door eigenaar, en gedocumenteerd. Alleen vereiste (kwaliteits)filters, en deze altijd gedocumenteerd/gevalideerd/geaccordeerd. Horizontale lineage vanaf de bronzone-ontkoppelviews, al dan niet via informatiegebieden. Verticale lineage door verwerking relevante delen FO in IDA/IGC.
Zoveel mogelijk RDBMS-onafhankelijk	Zoveel mogelijk gebruik maken van DataStage. SQL als tweede keus, geen PL/SQL of andere Oracle-specifieke functies (tenzij gegenereerd door DataStage).

Ontwerpuitgangspunt	**Impact**
Lineage mag niet gebroken worden	Alleen op DataStage gebaseerde transformaties tussen de input-gegevens (bronzone-ontkoppelviews / informatiegebieden) en de informatieproducten. Daardoor automatisch gegeneerde horizontale lineage. Gegevensvensters zijn hierop een uitzondering. Er wordt nog onderzocht of de lineage ook over de gegevensvenster-views heen ongebroken kan blijven.[100]
Minimale, en beheerbare, toolset	Oplossingen gebaseerd op DataStage (gegevensvensters op Oracle), tenzij duidelijk te argumenteren valt dat het daarmee niet op te lossen valt. Daarnaast bij de afdeling DWH reeds bekende (en door UWV toegestane) tools. Inzet IWS alleen voor opstarten master sequences, zodat de hiervoor benodigde kennis beperkt is.

[100] Zie paragraaf 3.11 (Lineage mag niet gebroken worden)

11 END-USER ZONE

De end-user zone valt buiten het DIM, en met uitzondering van een aantal BusinessObjects-universes, ook buiten het project DataFabriek.

De end-user zone wordt dan ook niet in dit Conceptueel ontwerp beschreven.

De eisen aan BusinessObjects-universes en de door die universes te gebruiken systeemaccounts zullen, als onderdeel van project-deliverable P187 (BO-standaarden) in een apart document worden vastgelegd.

N.B. PowerBI wordt binnen de end-user zone een steeds belangrijker BI-tool. UWV-brede inrichting van die tool valt buiten het project DataFabriek, en er is nog weinig over bekend.

12 METADATA-ZONE

De technische en functionele metadata betreffende de (verwerking van) gegevens in het DIM zijn terug te vinden in de **Metadata-zone**.

De **metadata repository** en de **informatiecatalogus** in deze zone bevatten alle technische respectievelijk functionele metadata die noodzakelijk is voor gegevensverwerking en -levering door het DIM. Deels wordt deze metadata door het DIM zelf gegenereerd (m.n. horizontale lineage), deels wordt deze overgenomen uit elders gecreëerde metadata (m.n. functionele gegevensdefinities brondata uit de FUGEMs). De koppeling tussen technische velden en hun functionele equivalent (de verticale lineage) zal ook binnen de DIM-metadata worden vastgelegd.

Metadata wordt, in het DIM, zowel gebruikt om beheerders en gebruikers te ondersteunen, als om geautomatiseerde processen te besturen en te volgen.

12.1 Typen metadata

Binnen het DIM zijn de volgende typen metadata relevant:

- Contract-metadata:
  - Interface-beschrijvingen
  - Leverafspraken
  - Communicatieafspraken en aanspreekpunten
- Functionele termen
- Lineage
  - Horizontaal
  - Verticaal
- Compliance-metadata
  - Bewijs rechtsgrond/proportionaliteit/subsidiariteit en/of doelbinding
  - Gebruiksbeperkingen
  - Vertrouwelijkheidsclassificaties
  - Bewaartermijnen
- Maskerings-metadata
- Kwaliteits-metadata
  - Meetcriteria voor datakwaliteit
  - De resultaten van de op basis daarvan uitgevoerde kwaliteitsmetingen
- Operationele metadata
  - Tracking- en controle-gegevens bronleveringen (“pakbon-metadata”)
  - Verwerkings-logs
  - Gebruiks-logs
- Technische metadata
- Omgevings-metadata

In de volgende paragrafen is elk van deze typen iets verder uitgewerkt. Meer detail in het document **Metadata binnen Datafabriek**.

12.1.1 Contract-metadata

Contract-metadata bestaat in het DIM voor leveringen van de bronnen aan het DIM, en voor informatieproducten die geleverd worden door het DIM.

In het eerste geval wordt de metadata gebruikt bij het verwerken van de bronleveringen, zowel door de geautomatiseerde processen (de harnassen van ontsluitingszone en bronzone), als, bij problemen, door DIM-beheer.

12.1.2 Functionele termen

Functionele termen beschrijven hoe binnen UWV tegen gegevenselementen aangekeken wordt. Functionele termen bevatten veelal functionele metadata, zoals definities, en gaan soms in op wat meer technischer zaken, zoals een formaat. Voor meer detail zie **Metadata binnen Datafabriek**.

Functionele termen worden gebruikt om gegevens uit de bronnen robuust en compliant te kunnen verwerken in het DIM, en deze gegevens zo eenduidig mogelijk (en weer compliant) middels informatieproducten te kunnen doorleveren aan de afnemers van het DIM.
Functionele termen zijn daarnaast van groot belang voor die afnemers bij het gebruiken van die informatieproducten.

Functionele termen worden opgenomen in het DIM wanneer deze erkend en herkend worden door de business. Vele functionele termen komen dan ook van buiten de Datafabriek en worden ingelezen vanuit de daarvoor gebruikte tooling. Zo worden functionele termen afkomstig uit FUGEM's en CGM ingelezen uit PowerDesigner. Functionele termen afkomstig uit de HUDSV worden ingelezen uit Sharepoint.
Ook zullen er functionele termen nodig zijn die middels de RLO aangeleverd zullen worden.

In alle gevallen is (de metadata-zone van) het DIM afnemer van de metadata en niet de leidende administratie.[101]
Alleen voor binnen het DIM gecreëerde (afgeleide) gegevens is het DIM de leidende administratie voor de bijbehorende gegevensdefinities. Ook in die gevallen blijft de business overigens nog steeds verantwoordelijk voor de functionele definities.

Het is van belang dat de afnemers en de bronnen hetzelfde begrip van de data en de metadata hebben. Daarom faciliteert het DIM ook in de koppeling tussen functionele termen uit de verschillende metadata-administraties binnen UWV. Denk bijvoorbeeld aan een functionele term uit de HUDSV, waar in de definitie van deze functionele term verwezen kan worden naar functionele termen afkomstig uit een FUGEM of het CGM. Deze termen worden dan aan elkaar gekoppeld in IGC. Ook voor deze koppelingen geldt dat het DIM afnemer is van deze metadata en niet de leidende administratie.

12.1.3 Lineage

De horizontale lineage volgt de gegevensstromen door het DIM, van bron tot afnemer.
Deze lineage wordt grotendeels automatisch gegenereerd door de ETL-tooling.

De verticale lineage legt het verband tussen een functionele term en een technische tabel of kolom. De verticale lineage maakt het samen met de (meer technische) horizontale lineage mogelijk om terug te traceren hoe data in een informatieproduct is ontstaan uit gegevens in de bron(nen). Andersom is het ook mogelijk om vanuit gegevens uit een bron te traceren binnen welke informatieproducten dit brongegeven terug te vinden is.

Bij generieke ETL-processen (zoals de harnassen) levert, juist door de generiekheid ervan, de automatisch gegenereerde horizontale lineage slechts beperkt inzicht.

Vandaar dat de lineage binnen het DIM pas begint bij de bronzone-ontkoppelviews.
Metadata-technisch is dat geen probleem:

- Tussen bronlevering en bronzone, en binnen de bronzone-ontkoppelviews, vinden alleen zuiver op de technische structuur gerichte, en zwaar gestandaardiseerde,

[101] De leidende administratie zal altijd getoond worden binnen het DIM om duidelijk te maken waar de functionele term vandaan komt.

gegevenstransformaties plaats; gegevens in de ontkoppelviews zijn dus vrijwel direct te relateren aan gegevens in de bron.
- Pas vanaf de bronzone-ontkoppelviews zijn de gegevens bruikbaar voor verdere verwerking en/of toegankelijk voor eindgebruikers

12.1.4 Compliance-metadata

Door de bronnen geleverde compliance-metadata wordt gebruikt om gegevens uit die bronnen compliant te kunnen verwerken in het DIM, en ook weer compliant (middels informatieproducten) te kunnen doorleveren aan de afnemers van het DIM.

Bij dat doorleveren wordt ook compliance-metadata van de afnemers gebruikt: hun bewijs dat ze rechtsgrond/proportionaliteit/subsidiariteit en/of doelbinding hebben wordt vastgelegd.

Bewaartermijnen worden, afhankelijk van het type data, zowel door bronnen als afnemers geleverd. Bewaartermijnen zijn de basis voor de archiverings- en vernietigingsfuncties van het DIM.

12.1.5 Maskerings-metadata

Als van een veld is aangegeven dat het identificerend is dan leggen we voor dat veld functionele en technische maskerings-metadata vast.

De functionele maskerings-metadata beschrijft de requirements met betrekking tot het maskeren. De technische maskerings-metadata beschrijft de technische oplossing rond een maskeringsklasse.

Functionele maskerings-metadata die opgeslagen dienen te worden in het DIM zijn:
- Een functionele omschrijving hoe te maskeren, inclusief koppelbaarheid, uniformeerregels en opsplitsing in (afgeleide) velden
- Een voorbeeld ter verduidelijking
- Een link naar alle kolommen in het DIM waar de data volgens deze maskeringsklasse is gemaskeerd

Technische maskerings-metadata (stuur-metadata) die opgeslagen dienen te worden in het DIM zijn:
- Indicator identificerendheid
- Maskeringsmethode (de naam van de bouwsteen en de koppeling naar de maskeringsklasse)
- Indicator afgeleid veld

De functionele maskerings-metadata zal opgeslagen worden in IGC en ook beschikbaar gemaakt worden middels de uit IGC gegenereerde Excel spreadsheet **Bescherming van persoonsgegevens in het DIM - Matrix Maskeringsklassen**. De spreadsheet is bedoeld voor stakeholders die wel de maskeringsmetadata moeten kunnen inzien, maar geen toegang hebben tot IGC.
Het voordeel van de functionele maskerings-metadata opslaan in IGC is dat een technische link gelegd kan worden tussen de maskeringsklasse en de technische tabellen en kolommen uit het DIM. Hiermee kan de maskerings-metadata dus onderdeel gemaakt worden van de lineage.
De technische maskerings-metadata zal opgeslagen worden als onderdeel van de stuur-metadata.
Voor meer detail zie het document **Datafabriek - Bescherming van persoonsgegevens in het DIM**.

12.1.6 Kwaliteits-metadata

Kwaliteits-metadata bestaat in twee vormen:

- Meetcriteria voor datakwaliteit
- De resultaten van de op basis daarvan uitgevoerde kwaliteitsmetingen

Meetcriteria zijn meestal afgeleiden van elders al binnen het DIM aanwezige metadata. Voorbeelden daarvan zijn of een veld verplicht is of niet (afgeleid uit de RLO), en of een gegeven voldoende actueel is (afgeleid uit log-metadata).

Het inrichten van de datakwaliteitstooling (QualityStage, IBM Information Analyzer) is geen onderdeel van de fasen 4 of 5.

12.1.7 Operationele metadata

Voor het DIM zijn vier vormen operationele metadata relevant:

- **Tracking- en controle-gegevens bronleveringen ("pakbon-metadata")**
  Deze metadata beschrijft een specifieke bronlevering.
  De metadata wordt gebruikt bij het verwerken/valideren van die bronlevering, en vervolgens gearchiveerd.
- **Verwerkings-logs**
  Een audit trail van alle binnen het DIM uitgevoerde geautomatiseerde processen.
  Met name relevant voor DIM-beheer.
- **Gebruiks-logs**
  Een audit trail van alle op het DIM uitgevoerde leesacties en wijzigingen (door afnemers, ontwikkelaars en beheerders DBA's, en door geautomatiseerde processen).
  Met name gebruikt voor operationeel risicobeheer en compliance-validatie.
- **Actualiteits-metadata**
  Op welke datum/tjjd een gegevensobject en/of informatieproduct voor het laatst is bijgewerkt, met daarbij bovendien, waar relevant, de actualiteit van de gebruikte input-gegevens.
  Actualiteits-metadata wordt vooral gebruikt bij het om de afhankelijkheden tussen bronleveringen en de informatiegebieden/informatieproducten die erop gebaseerd zijn te besturen.
  Actualiteits-metadata is daarnaast ook belangrijke meta-informatie voor de afnemers van informatieproducten.

12.1.8 Technische metadata

Technische metadata wordt binnen standaardpakketten gebruikt, en meestal opgeslagen in een (proprietary) repository.
Technische metadata is meestal alleen toegankelijk via de ontwikkel- en beheer-tools van het standaardpakket.

In de context van het DIM gaat het hier vooral om de definities van ETL-jobs (en de logica daarbinnen) in DataStage, en om de database-metadata van Oracle.

N.B. Vaak is het mogelijk (bij DataStage ook) om bij een OTAP-promotie de technische metadata te promoveren, in plaats van de, op basis van die technische metadata gegeneerde, executable code.

12.1.9 Omgevings-metadata

Om te voorkomen dat er bij promotie van releases (van O naar T, naar A, naar P) handmatige wijzigingen moeten worden doorgevoerd bevat elk van die OTAP-omgevingen omgevings-metadata (ook wel bekend als "environment variables").

Voorbeelden hiervan zijn:
- database connectie gegevens
- email-gegevens van DIM beheer
  (voor O/T vaak vervangen door die van de tester of ontwikkelaar)
- Linux-paden op de DataStage server

12.2 Metadata-componenten

In het DIM wordt metadata op meerdere manieren gebruikt, en in meerdere metadata-componenten opgeslagen en beheerd. Elk van deze componenten gebruikt en/of beheert één of meer metadata-typen.

De metadata-componenten zijn deels verzamelingen van documenten (meestal conform een vast sjabloon), deels kant-en-klare oplossingen (IBM-tooling), en deels maatwerk-oplossingen.

Metadata-componenten o.b.v. documenten:
- GLO[GL(19]/SNO Leveringscontracten (naar het DIM en uit het DIM)
- RLO (Record LayOut) Specificatie van de gegevensleveringen (door de bron) aan het DIM
- Functioneel Ontwerp Beschrijvingen van requirements en logische data flows

Metadata-componenten o.b.v. IBM Tooling:
- IDA/IGC/DataStage Functionele termen, tabellen en kolommen, maskeringsklassen (inclusief verticale en horizontale lineage)

Metadata-componenten als maatwerk-oplossing:
- Stuur-metadata Het sturen van ETL-processen
- Log-metadata Het loggen van ETL-processen

In de volgende paragrafen worden de componenten van de metadata-zone verder uitgediept, wordt duidelijk wie de doelgroep is en waar meer detail terug te vinden is.

12.2.1 Metadata-componenten o.b.v. documenten

Een aantal metadata-componenten zijn eigenlijk verzamelingen van documenten (meestal conform een vast sjabloon), opgeslagen in een vaste folder-structuur. Ze worden vooral gebruikt om afspraken tussen verschillende partijen en/of teams vast te leggen.

12.2.1.1 GIA en GLA~~GLO en SNO~~

De GegevensInwinningsAfspraak (GIA) ~~GegevensLeveringsOvereenkomst (GLO)~~ is het contract waarin de afspraken rondom een levering van gegevens door een bron **aan** het DIM zijn vastgelegd.
De GegevensLeveringsAfspraak (GLA) ~~ServiceNiveauOvereenkomst (SNO)~~ is het contract waarin de afspraken rondom een levering van gegevens aan een afnemer **uit** het DIM zijn vastgelegd.

Het GIA~~GLO~~-proces wordt, ten tijde van het schrijven van dit ontwerp, aangepast door het Implementatieteam GD/DWH (geen onderdeel van het project DataFabriek).
~~Onderdeel van die aanpassing is vervanging van GLO en SNO door nieuwe contractvormen, met een nieuwe naam.~~ Ook wordt daarbij gekeken in welke mate de Gegevensdiensten-applicatie UGC een rol kan spelen bij beheer/opslag van deze contracten.

GIA's ~~GLO's~~ en GLA's ~~SNO's~~ bevatten contract-metadata en compliance-metadata

12.2.1.2 RLO

Een RLO (Record LayOut) is onderdeel van de GIA~~GLO (GegevensLeverings Overeenkomst)~~, en bevat een volledige en gedetailleerde beschrijving van de inhoud van een bronlevering.

De RLO's zijn eigendom van (en worden beheerd door) de bron.

De RLO definieert die inhoud ook in functionele termen, door aan iedere tabel en iedere kolom uit de bronlevering een definitie te koppelen die erkend en herkend wordt door de business. Bij voorkeur gebeurt dit door een relatie te leggen tussen de gegevens in de levering en termen afkomstig uit een FUGEM en of het CGM. Het definiëren van de inhoud in functionele termen is randvoorwaardelijk voor het correct verwerken in informatieproducten. Dit is vooral van belang voor de afnemers.

De RLO beschrijft welke gegevens door de bron geleverd worden, hoe deze technisch geïnterpreteerd moeten worden. Denk hierbij aan veldformaten, sleutels en historisch gedrag. Deze metadata is onmisbaar voor correcte verwerking en opslag van de gegevens in het DIM. Dit is vooral van belang voor de technische ontwikkelteams binnen Datafabriek.

De RLO's zijn daarmee voor het DIM één van de belangrijkste bronnen van metadata; de stuur-metadata en de verticale lineage zijn er grotendeels op gebaseerd. Meer detail over de RLO en de invulling ervan is te vinden in de **Invulhulp RLO**.

RLO's bevatten contract-metadata, maskerings-metadata, functionele termen, (input voor de) verticale lineage, compliance-metadata

12.2.1.3 Functioneel Ontwerp

Het Functioneel Ontwerp beschrijft de requirements en de logische data flows die ten grondslag liggen aan de tabellen en kolommen, alsmede de ETL-jobs in het DIM. Concreet bestaat dit onder andere uit:

- Een beschrijving van de requirements
- (relevante delen van de) Logische Data Modellen (LDM) binnen de verschillende lagen van het DIM, i.e. Bronzone, Integratiezone en Bedrijfszone, inclusief
    - koppeling voor iedere tabel en kolom naar een functionele term afkomstig uit IGC
    - (te onderzoeken) mogelijke koppeling voor iedere tabel en kolom naar de maskeringsklasse
- Een logische mapping tussen de LDM's uit de verschillende lagen

Het Functioneel Ontwerp vormt dus de basis voor, en is input voor, het technische ontwikkelwerk. Voor meer detail over het maken van een Functioneel Ontwerp en de gebruikte standaarden, zie de documenten in H1.3 **Standaarden modelleren en FO voor het DIM** en **WoW Modelleren en FO maken in IDA**.

FO's bevatten functionele termen en (input voor de) horizontale en verticale lineage.

12.2.2 Metadata-componenten o.b.v. standaard-pakketten

Verschillende metadata-componenten zijn afkomstig uit de in de Europese Aanbesteding aangekochte IBM-tooling. In deze tools worden verschillende typen metadata vastgelegd. Doordat de tools uitstekend op elkaar aansluiten kunnen de verbanden tussen de verschillende typen metadata inzichtelijk gemaakt worden, resulterend in verticale- en horizontale lineage.

Daarnaast bevat de Oracle-database eigen technische metadata.
De IBM-tooling maakt hier deels weer gebruik van,

12.2.2.1 IDA

Infosphere Data Architect (IDA) is de modelleringstool binnen de IBM-suite. De tool wordt gebruikt om Logische Data Modellen (LDM) te bouwen en de logische mappings tussen deze LDM's inzichtelijk te maken.

Het LDM van de **bronzone** wordt automatisch gegenereerd in IDA op basis van de beschikbare metadata in de RLO. Het LDM voor de **integratiezone** en de **bedrijfszone** zijn maatwerk-oplossingen. Voor de LDM's van integratiezone en bedrijfszone wordt voor iedere entiteit en ieder attribuut de link gelegd naar een functionele term uit IGC, en mogelijk ook (nader te onderzoeken[GL(20]) naar de te hanteren maskeringsklasse. Hierdoor wordt de basis gelegd voor de verticale lineage.

In IDA zullen ook logische mappings gebouwd worden tussen
- (delen van het) LDM uit de Bronzone naar (delen van het) LDM uit de Integratiezone;
- (delen van het) LDM uit de Integratiezone naar (delen van het) LDM uit de Bedrijfszone.

Hierdoor wordt de basis gelegd voor de horizontale lineage.

De LDM's en de logische mappings in IDA kunnen automatisch beschikbaar gemaakt worden als onderdeel van het Functioneel Ontwerp. Via een Functioneel Ontwerp zijn deze modellen/mappings ook beschikbaar voor de eindgebruikers; zij kunnen de Logische Data Modellen en logische mappings immers niet raadplegen in IDA.

Voor meer detail over het gebruik van IDA en de gebruikte standaarden, zie de documenten **Standaarden modelleren en FO voor het DIM** en **WoW Modelleren en FO maken in IDA**.

IDA bevat functionele termen, en (input voor de) horizontale en verticale lineage.

12.2.2.2 IGC

Infosphere Information Governance Catalogue (IGC), ook onderdeel van de IBM-suite, is de plek waar de metadata samenkomt en aan elkaar gekoppeld wordt.

Deze component bevat:
- alle functionele termen afkomstig uit FUGEM, CGM, HUDSV en RLO
- alle tabellen en kolommen uit het DIM
- de koppeling tussen de functionele termen en de tabellen en kolommen, zoals geïmporteerd uit de Oracle-metadata (verticale lineage)
- de koppeling tussen de metadata van de tabellen en kolommen in de verschillende zones binnen het DIM, i.e. de bronzone, integratiezone, bedrijfszone (horizontale lineage)
- maskerings-metadata
  maskeringsklassen met daarin functionele maskerings-metadata
- *(te onderzoeken*[GL(21]*)* mogelijk de koppeling tussen de verschillende tabellen en kolommen uit het DIM en de maskeringsklassen

IGC wordt voor het raadplegen van bovenstaande metadata beschikbaar gesteld voor de eindgebruikers. Meer detail is terug te vinden in **Metadata binnen Datafabriek**.

IGC bevat functionele termen, horizontale en verticale lineage, compliance-metadata, maskerings-metadata, en in de toekomst ook kwaliteits-metadata.

12.2.2.3 DataStage

DataStage is de ETL-tool binnen de IBM-suite. M.b.v. DataStage gebouwde en uitgevoerde ETL-transformaties ("stages") zijn gedefinieerd middels (proprietary) technische metadata.

De ETL-transformaties worden gebouwd tussen de tabellen en kolommen in de fysieke (Physical) Data Modellen (PDM) van de diverse DIM-zones, en resulteren in "mappings" tussen de tabellen/kolommen in die PDM's.

Een PDM kan automatisch gegenereerd worden vanuit een LDM in IDA en middels automatisch gegenereerde DDL voor DataStage beschikbaar gemaakt worden.

Doordat de ETL-transformaties in DataStage gedefinieerd zijn als metadata, kunnen zowel de daaruitvolgende mappings als de PDM's uit de verschillende lagen van het DIM in IGC geladen worden, inclusief de connecties ertussen. Hierdoor wordt de horizontale lineage automatisch inzichtelijk gemaakt in IGC.

De technische metadata in DataStage zijn niet zichtbaar voor de eindgebruikers, en voor ontwikkelaars alleen indirect (via de Datastage-ontwikkeltools)

DataStage bevat technische metadata en (input voor) horizontale lineage.

12.2.2.4 QualityStage, IBM Information Analyzer

Hier nog niet beschreven; het inrichten van de datakwaliteitstooling (QualityStage, IBM Information Analyzer) is geen onderdeel van de fasen 4 of 5.

12.2.2.5 Oracle-metadata

De Oracle-metadata bevat de database-metadata die de tabellen en kolommen zoals deze in de database voorkomen beschrijven.

Concreet gaat het hier om tabellen en kolommen uit de verschillende lagen van het DIM, dus de bronzone, integratiezone en bedrijfszone, maar ook om de koppelvlakken daartussen, zoals de bronzone ontkoppelviews.

De metadata kunnen worden geïmporteerd in IGC, zodat ze als zogenaamde 'physical assets' in IGC beschikbaar komen. Hierdoor zijn de tabellen en kolommen daar koppelbaar in de horizontale en verticale lineage.

12.2.3 Metadata-component Stuur-metadata

Stuur-metadata wordt gebruikt door allerlei, meer of minder generieke, ETL-processen, en wordt gebruikt om de agility en robuustheid van die processen te verhogen.

De grootste gebruikers zijn de ETL-harnassen.

Andere processen die gebruik maken van stuur-metadata zijn:

- Datakwaliteitscontroles
- Processen voor archivering/vernietiging van gegevens

De stuur-metadata is niet beschikbaar voor eindgebruikers.[102]

[102] Zie (verwijzingen) voor de metadata-oplossingen t.b.v. eindgebruikers

Stuur-metadata wordt historisch vastgelegd. Hierdoor is het mogelijk te zien op welk moment er historische veranderingen hebben plaatsgevonden van bijv. de betekenis van gegevens binnen bronleveringen.

Stuur-metadata is geversioneerd door middel van een start- en een einddatum waardoor het ook mogelijk wordt om toekomstige wijzigingen alvast voor te bereiden en klaar te zetten voor gebruik.

De stuur-metadata bevat contract-metadata, compliance-metadata, maskerings-metadata, en in de toekomst ook kwaliteits-metadata.

De metadata-component voor stuur-metadata bestaat uit binnen UWV ontworpen en ontwikkelde gegevensstructuren in een Oracle-database, gecombineerd met ook binnen UWV ontwikkelde beheer-functionaliteit.
Maatwerk bleek noodzakelijk omdat de standaard IBM-tooling hiervoor geen handige oplossing bleek. Zie Bijlage C: Rationale maatwerk stuur-metadata voor de achtergrond van deze keuze voor maatwerk.

12.2.3.1 Categorieën stuur-metadata

De stuur-metadata valt uiteen in een aantal categorieën:

- **Stuur-metadata m.b.t. bronleveringen**
  Metadata t.b.v. de verwerking van bronleveringen naar de bronzone, waaronder:
  - Formaat en inhoud van de bronlevering
  - Contractafspraken voor de bronlevering (levermoment, aanspreekpunt, etc)
  - Op de bronlevering uit te voeren verwerkingen en controles
  - Voor de verwerking relevante bestands- en database-locaties

  Het doel van deze stuur-metadata is het sturen van de werking van de verschillende harnassen bij het verwerken van de bronlevering vanaf de ontsluitingszone tot en met de bronzone. Ook kunnen de ontkoppelviews op basis hiervan gegenereerd en onderhouden worden

- **Stuur-metadata m.b.t. informatiegebieden en -producten**
  Metadata voor het aanmaken/verversen van informatiegebieden en informatieproducten. Minder "rijk" dan de vorige categorie, omdat er minder generieke oplossingen worden toegepast.

- **Stuur-metadata m.b.t. Data Lifecycle Management**
  M.n. bewaartermijnen

N.B. Omdat er een duidelijke scheiding is tussen de verwerking van de bronleveringen naar de bronzone en de verwerking naar de integratie en bedrijfszone is de stuur-metadata hiervan ook gescheiden.

Zie Bijlage E: Overzicht stuur-metadata voor meer detail per categorie.

12.2.3.2 Technische opzet stuur-metadata

De stuur-metadata staat in een als maatwerk ontwikkeld schema van de Oracle database, en is dus géén onderdeel van de metadata binnen de diverse metadata-tools van de IBM Infosphere suite.

12.2.3.3 Beheer stuur-metadata

Er wordt een maatwerk-applicatie ontwikkeld voor het beheren van de stuur-metadata, zodat onderhoud en versiebeheer van de stuur-metadata stabiel kan plaatsvinden.

De applicatie zal bij het invoeren en opslaan controles uitvoeren zodat de kwaliteit van de stuur-metadata en, bijvoorbeeld, de werking van de harnassen gewaarborgd is. Verder zal de applicatie het release management van de stuur-metadata ondersteunen.

Voor meer details wordt verwezen naar de documentatie welke opgesteld is/wordt tijdens de realisatie.

12.2.4 **Metadata-component Log-metadata**

De metadata-component voor log-metadata bestaat deels uit standaard IBM-functionaliteit, maar daarnaast ook uit binnen UWV ontworpen en ontwikkelde gegevensstructuren in een Oracle-database, gecombineerd met ook binnen UWV ontwikkelde dashboards.

12.2.4.1 Standaard verwerkings-logging

De ETL-tooling biedt, ”van de plank” logging van ETL-jobs. De kracht van deze logging is dat deze geen verdere eisen stelt aan de ETL-jobs, en “altijd werkt” (mocht er onverhoopt een master sequence hard omvallen zonder naar de maatwerk-verwerkingslogs te kunnen schrijven, dan kan er nog altijd gebruik gemaakt worden van de standaard verwerkings-logs), de zwakte is dat de logging, door de generiekheid, minder geschikt is voor meer gericht gebruik, zoals voortgangsbewaking en trend-analyse van verwerkingsperformance.

Ook bij metadata-gedreven functionaliteit is de standaard-logging soms te weinig specifiek.

12.2.4.2 Maatwerk verwerkings-logs, algemeen

In aanvulling op de standaard geleverde basale logging bewaart het DIM, in maatwerk Oracle-datastructuren, ook additionele logging.
Deze logging wordt gecreëerd door in elke stap of module binnen een ETL-job één of meer ”logging-bouwstenen” op te nemen. Op deze wijze hebben we wel op het globale niveau van ETL-jobs informatie voor analyse-doeleinden en voortgangsbewaking.

Het gaat hier om zaken als:

- Start tijdstamp
- Eind tijdstamp
- Status
- Meldingen

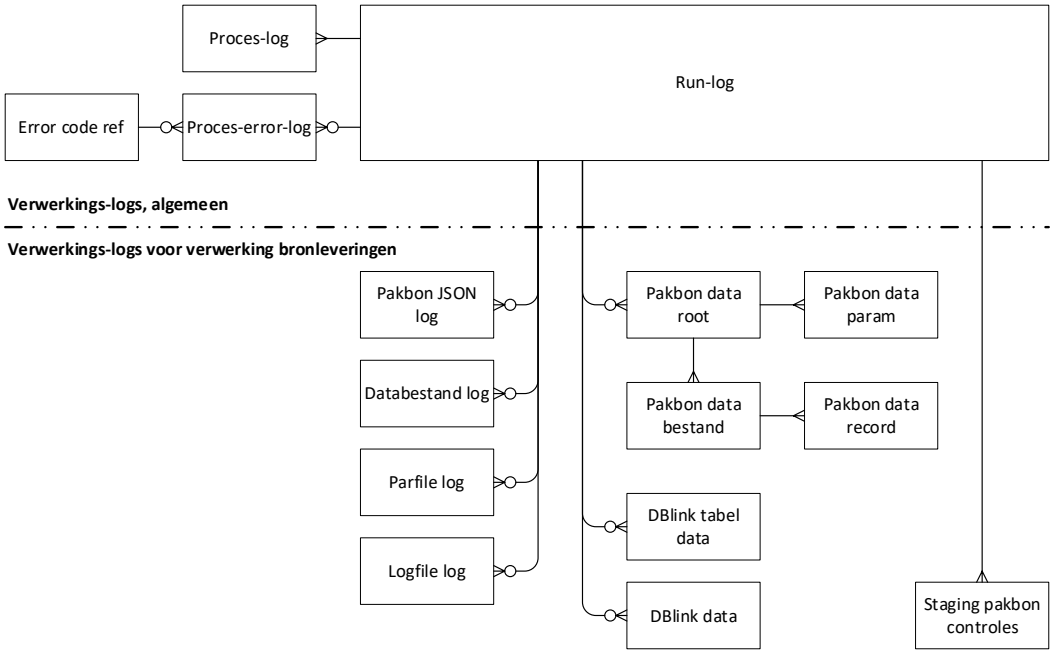
N.B. Maatwerk-logging is bedoeld als uitbreiding op de standaard-logging, niet als vervanging ervan. Conform het ontwerpuitgangspunt “Minimale, en beheerbare, toolset” geldt ook hier: “als het al bestaat, dan bouwen we het niet".

12.2.4.3 Maatwerk verwerkings-logs voor verwerking bronleveringen

Om de verwerking van bronleveringen ook in detail te kunnen monitoren en om probleemoplossing te faciliteren worden alle stappen in de ETL-harnassen (en eventueel daarvan afgeleid ETL-maatwerk) gelogd.
Net als bij de algemene verwerkings-logs wordt deze log opgeslagen in maatwerk Oracle-datastructuren.

Door deze logging is het ook mogelijk om, voor ETL-harnassen, gedetailleerde foutanalyses uit te voeren, en/of performance-trends te volgen en trend-rapportages te maken voor de performance van de ETL-harnassen.

N.B. Afgekeurde gegevensregels uit een bronlevering worden in een reject-bestand weggeschreven, en niet in de log-metadata. Dit omdat het zonder melding leveren van, bijvoorbeeld, een extra attribuut anders zou resulteren in een foutmelding voor elke regel, en dus zou kunnen leiden tot het vollopen van de meldingen-tabel. Daar dit niet wenselijk is hebben we besloten dit type afkeuringsmelding in een bestand te zetten. Dit maakt eventuele communicatie met de bron over afgekeurde gegevensregels ook makkelijker.



(dit plaatje is een schets; uitwerking in apart document)

N.B. Dit extra stuk maatwerk-logging is bedoeld als uitbreiding op de algemene maatwerk-logging, niet als vervanging ervan. Het is dus belangrijk om algemene onderdelen en bronverwerkings-specifieke onderdelen zo veel mogelijk gescheiden te houden, maar wél zo in te richten dat ze, bij monitoring en analyse, als één geheel kunnen worden gezien. E.e.a. conform het ontwerpuitgangspunt "Specifiek waar mogelijk, gemeenschappelijk waar noodzakelijk of waardevol".

12.2.4.4 Maatwerk verwerkings-logs voor overige verwerkingen

Op dit moment nog niet bekend/relevant. Zullen aan dit document worden toegevoegd zodra dit relevant is.

12.2.4.5 Dashboards op log-metadata

Waar mogelijk worden de standaard-dashboards van DataStage gebruikt. Deze dekken echter alleen de basale verwerkingslogs.

Extra dashboards op de maatwerk-verwerkingslogs worden tijdelijk gebouwd met BusinessObjects.

Ontwerp en implementatie van de "Prodmon-opvolger" is onderdeel van Fase 5, en nog niet beschreven in deze versie van het conceptueel ontwerp.

12.2.5 Metadata-component Actualiteits-metadata

In deze component wordt:

- voor elke bronlevering (en elke tabel daarbinnen) bijgehouden tot wanneer deze verwerkt is in de ontkoppelviews van de bronzone
- voor elk informatiegebied bijgehouden wanneer deze voor het laatst ververst is
- voor elke datamart en elke gematerialiseerde gegevensvensterview bijgehouden wanneer deze voor het laatst ververst is
- voor tijd-gedreven informatieproducten bijgehouden wat de actualiteit was van de gebruikte gegevens.

Actualiteits-metadata wordt vooral gebruikt om de afhankelijkheden tussen bronleveringen en de informatiegebieden/informatieproducten die erop gebaseerd zijn te besturen. Hiervoor hebben alleen ETL-processen toegang tot de component nodig.
De metadata-component voor actualiteits-metadata bestaat uit binnen UWV ontworpen en ontwikkelde gegevensstructuren in een Oracle-database.

N.B. Actualiteits-metadata is ook van belang voor de afnemers van informatieproducten. Waar cruciaal kan deze metadata in het informatieproduct zelf worden opgenomen.

Meer algemene ontsluiting tot de actualiteits-metadata (bijvoorbeeld middels dashboards) is (nog) niet in scope van het project DataFabriek.

N.B. Actualiteits-metadata lijkt op log-metadata, maar is net iets anders; de log-metadata is primair voor beheerders, en beschrijft de **status** van de **processen**, de actualiteits-metadata is voor de afnemers (initieel alleen ETL-processen, later ook "echte" afnemers), en beschrijft de status van de **data**.

13 BEHEERZONE

De beheerzone is een "verzamelbak" van alle functionaliteit die nodig is om het DIM als geheel te kunnen beheren.

Deze functionaliteit valt uiteen in de volgende categorieën:

- Scheduling
- Leveringsbeheer
- Toegangsbeheer
- Logging & monitoring (processen)
- Logging en monitoring (gebruikers en IV-medewerkers)
- DBA-tooling
- Encryptie / tokenisation

13.1 Scheduling

De scheduling van het DIM is gebaseerd op IBM Workload Scheduler (IWS).

Deze scheduler wordt vooral gebruikt om de "hartslag" van het DIM te leveren; de dag/week/maand/jaar-schema's waarin de bronnen leveren en waarin de informatieproducten geleverd worden.

Deze IWS-scheduling bevat zo min mogelijk intelligentie.

Dit wordt op drie manieren bereikt:

- **De scheduler start steeds hele ETL-ketens.**
  De stappen binnen die keten zijn weer vastgelegd in, binnen DataStage gebouwde, Master Sequences[103].

- **De scheduler wordt alleen gedreven door de "hartslag".**
  Gedetailleerdere triggers (bijvoorbeeld het arriveren van een bronlevering in een inbox) worden binnen de master sequences afgehandeld.
  Er zijn drie gevallen waarin deze hartslag niet voldoet:
  - Bij het inlopen van achterstanden moet het hart "even sneller slaan" om een aantal opeenvolgende bronleveringen (van dezelfde bron) direct achter elkaar te kunnen verwerken.
  - Voor onregelmatige bronleveringen wordt de DIM-scheduler waarschijnlijk getriggerd via enterprise scheduling (zie volgende paragraaf).
  - Voor ad hoc leveringen binnen een interactief proces (bijvoorbeeld t.b.v. traceerbare reparaties, zie Bijlage G: Overschrijven kan niet, verbeteren wel) wordt de DIM-scheduler óf niet gebruikt, of handmatig getriggerd.

  De precieze aanpak voor deze gevallen zal pas worden ontworpen als er een benoemd requirement op dit gebied is.

- **Ontkoppeling i.p.v. dependency management**
  Er is, bijvoorbeeld, niet één keten van de verwerking van een bronlevering tot en met de (deels) erop gebaseerde informatieproducten. De veel-op-veel relatie tussen bron en informatieproduct zou dan namelijk zeer complex worden.
  In plaats daarvan is er een aparte groep ketens (één per bron) voor de verwerking van een

[103] Zie paragraaf 8.6 (Master Sequence bronverwerking~~Master Sequence bron-verwerking~~) en 9.4~~9.6~~ (Master sequence informatievoorziening)

bronlevering tot en met de bronzone, en een tweede groep ketens (één per informatiegebied) voor de verwerking van de bronzone-data in integratiezone en informatieproducten. Deze tweede groep ketens wordt via IWS "gewoon" op basis van een hartslag gestart, controleert zelf, binnen de Master Sequence, of alle ervoor benodigde brondata beschikbaar is, en wacht zo nodig tot dat wel zo is.

Deze ontkoppeling heeft als bijkomend voordeel dat het herstarten van ETL-jobs eenvoudiger is; er hoeft immers geen rekening gehouden te worden met de, elders bewaakte, dependencies.

N.B. Het DIM gebruikt géén scheduling t.b.v. load management; dat wordt overgelaten aan de daarvoor bedoelde tools op OS- en database-niveau, en binnen de ETL-engine.

13.1.1 Enterprise scheduling

IWS is ook de standaard-tool voor applicatie-overstijgende scheduling. Inzet als zodanig valt buiten het DIM, en ook buiten DIM-beheer.

Als een schedule binnen het DIM moet worden opgestart door een applicatie-overstijgende schedule, dan verdient het de voorkeur dat dit zo ontkoppeld mogelijk gebeurt.

De precieze aanpak van dergelijke twee-laags-scheduling zal pas worden ontworpen als er een benoemd requirement op dit gebied is.

13.2 Leveringsbeheer

Het contractmanagement van de leveringen: door de bronnen aan het DIM, en door het DIM aan de afnemers. Het DIM sluit hier waarschijnlijk aan op UGC; de UWV-standaard voor dergelijke contracten.

Inrichting van het contract management wordt, buiten het project, uitgevoerd door de lijn. Leveringsbeheer wordt in dit conceptueel ontwerp dus niet verder beschreven.

13.3 Toegangsbeheer

De beveiliging van de toegang tot de gegevens in het DIM ligt "dichter bij de database" (en dus niet vooral in BI-tooling). Dit maakt het toegangsbeheer complexer.

De verwachting is dat er hulpmiddelen nodig gaan zijn om DIM-beheer hierbij te ondersteunen. Voornaamste startpunt daarvoor is de wijze waarop BusinessObjects op het DIM gaat worden aangesloten (deliverable van fase 4). Ontwikkeling [GL(22]*van eventuele tooling valt onder fase 5, en is in deze versie van het conceptueel ontwerp daarom nog niet beschreven.*

13.4 Logging & monitoring (processen)

Logging van de diverse geautomatiseerde processen gebeurt deels middels standaardfunctionaliteit van DataStage, en deels in de log-metadata[104], waarbij gebruik van die standaardfunctionaliteit, zoals altijd, de voorkeur heeft.

Monitoring gebeurt deels m.b.v. standaardfunctionaliteit van DataStage, deels op basis van door de ETL-processen verstuurde alerts, en deels door een DIM-equivalent van ProdMon.[105]

[104] Zie paragraaf 12.2.4~~12.6~~ (Metadata-component Log-metadata~~Log-metadata~~)

[105] Het DIM wordt niet opgenomen in Prodmon; die tool wordt, samen met de legacy DWH's, uitgefaseerd

Ontwerp en implementatie [GL(23]*van de "Prodmon-opvolger" is onderdeel van Fase 5, en nog niet beschreven in deze versie van het conceptueel ontwerp.*

13.5 Logging & monitoring (gebruikers en IV-medewerkers)

Vanwege security en compliance worden zowel de A- als de P-omgeving gemonitord.
Dit gebeurt allereerst conform de UWV-standaard (logs doorsturen naar Q-Radar, log-analyse centraal).

Daarnaast wordt middels eigen, door de DBA's ontwikkelde, scripts gecontroleerd:

a) of de autorisaties op de database kloppen met die in ABS
b) of er afnemers zijn met toegang tot zowel gemaskeerde als ongemaskeerde data
c) of de bronzone wordt benaderd (of: kan worden benaderd) buiten de ontkoppelviews om
d) of de DIM-gegevens worden benaderd (of: kunnen worden benaderd) buiten de informatieproducten in de bedrijfszone om
e) of er extreem belastende query's en/of gebruikers zijn

Controles c, en d gelden daarbij ook voor systeemaccounts, zoals gebruikt door ETL-processen en BI-tools.

Ook bovenstaande gebeurt zowel op de A- als de P-omgeving. Log-analyse (of verwerking van alerts) valt onder de taken van de DBA's en/of DIM-Beheer.

De DBA's controleren ook hun eigen IV-collega's (weer middels door henzelf ontwikkelde scripts):

a) of de A- en P-omgevingen alleen door ontwikkelaars/beheerders benaderd (kunnen) worden na goedkeuring (bv. voor GAT-ondersteuning, of voor het oplossen van een calamiteit)[106]
b) of toegang tot A en P na afronden GAT-ondersteuning of oplossen van de calamiteit weer is afgesloten
c) of de ontwikkelaars/beheerders zich houden aan de afspraken m.b.t. versiebeheer, releasemanagement en OTAP-promoties

Tenslotte controleert de rekencentrum-provider de DBA's, en eventuele andere beheerders met rechten die standaard alleen aan rekencentrum-medewerkers zijn voorbehouden.
Bij IBM is daarvoor weinig ingericht (in plaats daarvan zijn er "fenced VLAN's" geplaatst tussen de DIM-omgeving en de rest van het UWV), bij DXC zal hiervoor waarschijnlijk [GL(24]hun PAM-oplossing[107] worden ingezet.

13.6 DBA-tooling

Een verzamelcategorie voor alle functionaliteit die de DBA's van de afdeling DWH nodig hebben om de databases van het DIM te kunnen beheren. Deels zit deze functionaliteit in standaard Oracle-tools, deels zijn dit door de DBA's zelf geschreven scripts, bijvoorbeeld om de vullingsgraad van tablespaces in de gaten te houden.

Het DIM gebruikt dezelfde Oracle-tools als de bestaande DWH's, en op dezelfde wijze. Het gebruik ervan is in dit ontwerp dus niet verder beschreven.

Beheer-scripts e.d. worden door de DBA's zelf ontwikkeld, meestal als nieuwe versie van al bij andere DWH's in gebruik zijnde scripts. Ook deze scripts worden in dit ontwerp niet verder beschreven.

---

[106] Hoofdstuk 18 (Informatiebeveiliging en –beheer) beschrijft de "rode envelop" procedure die voor tijdelijke toegang gebruikt wordt.

[107] PAM: Privileged Account Management

13.7 Encryptie / tokenisation

Binnen het DIM zijn drie vormen van encryptie/tokenisation herkenbaar:

- Maskering van productie-data binnen het DIM
- Maskeringsdiensten t.b.v. analyseomgevingen buiten het DIM
- Anonimisering van test-data

13.7.1 Maskering van productie-data binnen het DIM

Deze vorm van maskering vervult een belangrijke rol bij het AVG-compliant maken van het DIM, en is in hoofdstukken 5 (Maskering en DLM) en 8 (Bronzone en ontkoppelviews) uitgebreid beschreven.

De gebruikte technologie is die van OPTIM, ingezet conform de "best practices" m.b.t. gegevensbeveiliging.

N.B. Aangezien deze maskering onderdeel is van de P-versie van het DIM komt dezelfde maskering ook voor in O, T en A. De gemaskeerde data is daarbij overigens niet tussen deze omgevingen uitwisselbaar, vanwege omgeving-specifieke geheime "salt's".

13.7.2 Maskeringsdiensten t.b.v. analyseomgevingen buiten het DIM

Als gemaskeerde analyseomgevingen buiten het DIM gegevens van elders willen koppelen met gegevens uit het DIM, dan kan dat op twee manieren:

1. De analyseomgeving krijgt **ongemaskeerde** gegevens uit het DIM, en kan deze dus zelf eenvoudig koppelen met gegevens van elders, en vervolgens **gemaskeerd** ter beschikking stellen aan rapportage/analyse-processen.
   Maskering vindt hier dus plaats buiten het DIM, en ook niet met de maskerings-functionaliteit van het DIM.
   Dit is alleen mogelijk als levering van ongemaskeerde gegevens aan die analyse-omgeving "rechtsgrond" heeft.
2. De analyseomgeving **stuurt de extra gegevens naar het DIM** om ze daar te laten maskeren, of laat er **door het DIM geleverde maskeringsbouwstenen** op los.
   Omdat DIM-data en extra gegevens nu op dezelfde wijze gemaskeerd zijn, zijn ze, binnen de analyseomgeving, nu ook in gemaskeerde vorm koppelbaar.

Ad 1)
Naar verwachting wordt deze benadering tijdelijk toegepast voor DMAP. Dit vereist wel expliciete instemming van BSO en FG; strikt genomen is opslag/gebruik van ongemaskeerde data slechts voor een beperkt deel van DMAP proportioneel.

Ad 2)
Dit zou, bijvoorbeeld, van toepassing kunnen zijn op de gegevens t.b.v. de statistici van SBK, die op BSN-niveau moeten kunnen koppelen met CBS-data.

Hoe deze dienst precies moet worden ingericht wordt, bij gebleken behoefte, in een latere projectfase bepaald. Aandachtspunt daarbij is dat het buiten het DIM ter beschikking stellen van DIM-maskering het risico op ongeoorloofd gegevensgebruik (w.o. ook "ontmaskering") door het combineren van gemaskeerde en ongemaskeerde gegevens vergroot.

13.7.3 Anonimisering van test-data

Conform UWV-beleid moet test-data (in O en T) anoniem zijn.

Het DIM sluit voor de anonimisering van deze data aan op de standaardprocessen en –tooling van het TSC (inclusief de DWH-specifieke uitbreidingen daarop). Maskering van test-data gebeurt dus niet binnen het DIM, of met de binnen het DIM gebouwde maskeringsbouwstenen.

N.B. De testgegevens in A zijn **niet** geanonimiseerd, aangezien een GAT realistische data vereist.
In plaats daarvan bevat A een (iets verouderde) kopie van P. Dit vereist overigens wel dat de toegang tot A net zo goed is afgeschermd als die tot P.

14 ARCHIVERING EN VERNIETIGING

Archivering en vernietiging gebeurt in het DIM op vele plaatsen en vele manieren:

- door bestanden eerst te verplaatsen naar archief-locaties, en op een later tijdstip te vernietigen;
- door tijdelijke database-opslag regulier te schonen;
- door Data Lifecycle Management (DLM) uit te voeren voor de kern-zones van het DIM, en hier ook definitieve verwijdering onderdeel van te laten zijn;
- door ook voor backups bewaartermijnen in acht te nemen;
- door het maken van procedurele afspraken waar applicatieve oplossingen niet mogelijk en/of haalbaar zijn

Zie hoofdstuk 5 (Maskering en DLM~~Maskering en DLM~~) voor meer informatie over het DLM binnen het DIM.

14.1 Archivering van bronleveringen

We onderscheiden op basis van de interface-standaarden 4 verschillende leveringstechnieken voor bronleveringen: bronbestanden, exports, database-links en berichten.

Voor al deze typen leveringen zal bij archivering sprake zijn van hot en cold storage. Bij hot storage zal de betreffende levering nog direct benaderbaar zijn. Bij cold storage zal de bronlevering terug gezet moeten worden naar hot storage.

De bronleveringen worden in hot of cold storage bewaard voor zolang het nodig en zinnig is om reparaties te kunnen doen. Ook kan het archief, indien noodzakelijk, gebruikt worden voor traceerbaarheid.

Voor elk bestand geldt een maximale bewaartermijn. Zodra deze bewaartermijn gepasseerd is zal het gearchiveerde bestand vernietigd worden.

De hot storage bewaartermijn zal in de orde van 2 maanden zijn (net iets meer dan de maximale periode tussen twee bronleveringen). De cold storage bewaartermijn is vele malen langer, en sluit aan op de algemene DWH-standaarden voor bestandsarchivering.[108]

Splitsing van de archieven in hot en cold storage is primair gericht op het versnellen van Disaster Recovery; vaak voldoet de inhoud van het hot archive (naast de database-backups en -logfiles) namelijk voor het weer consistent maken van de bronzone na een infra-calamiteit.

N.B. Disaster Recovery ondersteunende functionaliteit (zoals hot/cold storage) valt wél in scope van het project DataFabriek, maar de Disaster Recovery plannen/processen die er gebruik van maken niet. Die plannen/processen zijn dus ook niet beschreven in dit document. Zie ook paragraaf 18.3 (Business Continuity Management (BCM)~~Business Continuity Management (BCM)~~).

14.1.1 Archivering bronleveringen in archief-folder

Nadat een bronlevering ingelezen is in de **staging-laag** worden de geleverde bestanden (platte bestanden + pakbon of exports + par- en log-file), door de **laadlogica staging** (zie paragraaf 7.4), verplaatst naar een **archief-folder** voor de betreffende bron.

[108] Mochten deze algemene standaarden om compliance-redenen (AVG, Archiefwet) worden aangepast, dan zullen deze aanpassingen dus ook gelden voor het DIM.

Ook niet verwerkte (afgewezen) leveringen komen, inclusief eventuele reject-bestanden in die archief-folder terecht.
Database links worden gebruikt om rechtstreeks gegevens uit de bron in de staging-laag te kopiëren. Na het kopiëren zal er een tabeldump gedaan worden. Hierna geldt weer hetzelfde als voor bestanden.

De archief-folder is opgeslagen in **hot storage** en fungeert als **hot archive**, oftewel een archief waaruit opgeslagen bestanden met minimale inspanning terug kunnen worden gehaald om, in geval van infra-calamiteiten en/of andere reparaties vereisende issues, opnieuw verwerkt te kunnen worden.

14.1.2 Verplaatsing bronleveringen van archief-folder naar cold archive

De oudere bestanden in de **archief-folder** worden, als de kans dat ze voor reparatie en/of traceerbaarheid noodzakelijk zijn klein is geworden, verplaatst naar het **cold archive** (een archief op **cold storage**).

Dit gebeurt middels een apart, bij voorkeur metadata-gedreven, proces (waarschijnlijk een OS-script), dat één keer per maand draait.

Het moment van verplaatsen zal samenhangen met de frequentie van de levering van de bronbestanden. Exacte verplaatsingsmomenten zullen per bestand worden bepaald en liggen buiten de scope van dit document.

14.1.3 Vernietiging bronleveringen in cold archive

Bestanden in het **cold archive** waarvoor de bewaartermijn is verstreken worden vernietigd.

Dit gebeurt middels een apart, bij voorkeur metadata-gedreven, proces, dat één keer per jaar draait.

Dit proces verloopt in drie stappen:

1. Opstellen lijst te vernietigen bestanden
2. Goedkeuring vernietiging en archivering van de lijst (of een samenvatting ervan) in het Electronisch archief (handmatig proces, conform UWV-standaarden op dit gebied)
3. Vernietigen bestanden uit de lijst

De cold storage bewaartermijn sluit aan op de algemene DWH-standaarden voor bestandsarchivering. Exacte bewaartermijnen zullen per bestand worden bepaald en liggen buiten de scope van dit document. Het ontwikkelen van het vernietigingsproces valt óf in een latere projectfase, óf valt onder lijnverantwoordelijkheid.

14.2 Schoning van de staging-laag

Gegevens in de staging-laag zijn per definitie niet persistent. De gegevens zullen of direct nadat de gegevens in de bronzone zijn overgenomen of uiterlijk als eerste stap in de verwerking van een nieuwe bronlevering worden verwijderd.

Zie paragraaf 7.4 (Laadlogica staging) voor meer details.

14.3 DLM in de bronzone

De bronzone bevat drie groepen [GL(25]gegevens:

- Ongemaskeerde identificerende gegevens
- Gemaskeerde identificerende gegevens
- Niet-identificerende gegevens

N.B. De term "identificerend" moet hierbij vanuit het perspectief van de burger gezien worden. Identificerende gegevens betreffende UWV-medewerkers (in hun rol als medewerker, dus niet in hun rol als burger) worden in de bronzone niet gemaskeerd.

Middels de ontkoppelviews worden deze drie groepen [GL(26)]"virtueel" omgevormd tot drie andere:

- Ongemaskeerde gegevens
  (Ongemaskeerde identificerende gegevens + Niet-identificerende gegevens)
- Gemaskeerde gegevens
  (Gemaskeerde identificerende gegevens + Niet-identificerende gegevens)
- Niet-persoonsgegevens
  (uitsluitend Niet-identificerende gegevens; alleen relevant als de brondata zelf al geen persoonsgegeven was)

Ongemaskeerde gegevens in de bronzone vallen onder de bewaartermijnen voor operationele processen, gemaskeerde gegevens de (langere) bewaartermijnen voor rapportage/analyse-processen.

Binnen het DIM loopt het DLM van de **ongemaskeerde** gegevens, en dus ook de verwijdering daarvan, gelijk met het DLM van diezelfde gegevens binnen de bron.

N.B. Het DIM kent voor deze ongemaskeerde gegevens géén eigen archieffunctie, aangezien het archiveren van operationele gegevens de verantwoordelijkheid is van de bron, en daar dus ook al wordt uitgevoerd.

Het DLM van de **gemaskeerde** gegevens kent een eigen dynamiek, aangezien hier niet alleen het bron-DLM moet worden meegenomen, maar ook het gebruik van de gegevens binnen rapportage/analyse-processen. Ook geldt dat het DIM hier, in tegenstelling tot voor de ongemaskeerde gegevens, wél een archieffunctie heeft.

N.B. De bewaartermijnen voor klantgerichte processen (en dus ongemaskeerde gegevens) zijn nooit langer dan die voor analyse/rapportage-processen (en dus gemaskeerde gegevens). Vandaar dat voor de verwijdering van de niet-identificerende gegevens de bewaartermijnen voor analyse/rapportage kunnen worden aangehouden.

14.3.1 Verwijdering ongemaskeerde gegevens

In dit proces worden niet alle ongemaskeerde gegevens verwijderd, maar alleen het identificerende deel ervan (zie paragraaf hierboven).

Binnen het DIM loopt het DLM van deze **ongemaskeerde identificerende gegevens** gelijk met het DLM van diezelfde gegevens binnen de bron, hetzij door alle door de bron gemelde verwijderingen ook in het DIM uit te voeren, hetzij door de verwijderlogica van de bron in het DIM "na te spelen".

N.B. Deze verwijderingen kennen geen aparte goedkeurings- of vastleggingsstap, aangezien het DIM voor operationele gegevens geen formele archief-rol heeft. Wel wordt binnen het DIM een (in de stuur-metadata vastgelegde, bron-specifieke) "grace periode" aangehouden, waarbinnen de verwijderde gegevens niet meer zichtbaar zijn in de ontkoppel-views (en dus voor DIM-afnemers, maar nog niet fysiek verwijderd zijn. Dit om te zorgen dat leverfouten vanuit de bron niet resulteren in onomkeerbare schade aan de gegevens in het DIM.[109]

[109] Zie paragraaf 8.3.3 (Volgen van het DLM van de bron: harde en zachte verwijderingen)

14.3.2 Verwijdering gemaskeerde gegevens

In dit proces worden niet alleen de gemaskeerde gegevens verwijderd, maar ook de niet-identificerende gegevens.

Het DLM van de **gemaskeerde** gegevens kent een eigen dynamiek, aangezien hier niet alleen het bron-DLM moet worden meegenomen, maar ook het gebruik van de gegevens binnen rapportage/analyse-processen. Ook geldt dat het DIM hier, in tegenstelling tot voor de ongemaskeerde gegevens, wél een archieffunctie heeft.

Verwijdering van de gemaskeerde gegevens wordt daarom periodiek door een combinatie van stuur-metadata gedreven DIM-processen en handmatige stappen uitgevoerd, als volgt:

1. Een automatisch DIM-proces markeert rijen (versies) in de bronzone als "te vernietigen" o.b.v. algemene regels m.b.t. bewaartermijnen
2. De DIM-afnemers valideren (bijvoorbeeld m.b.v. BI-tools) dat deze gegevens inderdaad vernietigd kunnen worden.
3. Middels scripts worden alle gegevens waarvoor dit het geval is gemarkeerd met "vernietiging akkoord". Voor gegevens die, om welke reden dan ook, toch nog niet verwijderd mogen worden de vernietigingsmarkering vervangen door "vernietiging uitgesteld".[110]
4. De definitieve vernietigingslijst wordt opgesteld (door FB-DIM)
5. Na goedkeuring vernietiging (door afnemers en bronnen) wordt deze definitieve lijst (of een samenvatting ervan) opgeslagen in het Elektronisch archief (handmatig proces, conform UWV-standaarden op dit gebied)
6. Een automatisch DIM-proces verwijdert de met "vernietiging akkoord" gemarkeerde rijen uit de bronzone.

Periodiek (vermoedelijk jaarlijks) wordt voor de met "vernietigen uitgesteld" gemarkeerde gegevens, gecontroleerd of deze nog steeds bewaard dienen te blijven. Is dat inderdaad zo, dan blijft de markering staan, is dat niet zo, dan wordt de markering (middels scripts) weer teruggezet naar "te vernietigen" zodat de stappen 2-6 hierboven weer kunnen worden doorlopen.

14.3.3 Verwijdering niet-persoonsgegevens

Voor gegevens die geen betrekking hebben op burgers bevat de bronzone geen als "identificerend" opgeslagen gegevens, maar alleen niet-identificerende gegevens.
Op deze gegevens wordt het verwijderproces voor gemaskeerde gegevens (zie hierboven) toegepast. De bewaartermijnen binnen het DIM zijn dus die voor rapportage/analyse-processen.

14.4 DLM in de integratiezone

De integratiezone bevat vier categorieën gegevens:

a) ongemaskeerde persoonsgegevens
b) gemaskeerde persoonsgegevens
c) overige gegevens (überhaupt geen betrekking hebbend op personen)
d) door aggregatie anoniem geworden gegevens (eigenlijk een bijzondere vorm van "overige gegevens")

N.B. De term "persoon" moet hierbij vanuit het perspectief van de burger gezien worden. Gegevens betreffende UWV-medewerkers (in hun rol als medewerker, dus niet in hun rol als burger) kunnen in alle drie de groepen gegevens voorkomen.

[110] Deze expliciete markering zorgt ervoor dat die gegevens niet bij de eerstvolgende run van het proces onder punt 1 weer opnieuw worden gemarkeerd als "te vernietigen"

Gegevens in categorie (a) vallen onder de bewaartermijnen voor operationele processen, die in de overige categorieën onder de (langere) bewaartermijnen voor rapportage/analyse-processen.
De AVG is alleen van toepassing op de informatiegebieden in categorie (a) en (b).

Binnen het DIM loopt het DLM van **ongemaskeerde** persoonsgegevens, en dus ook de verwijdering daarvan, gelijk met het DLM van diezelfde gegevens binnen de bron. Aangezien dit ook al geldt voor het DLM van de bronzone, kan voor ongemaskeerde informatiegebieden in de integratiezone worden volstaan met het volgen van het DLM van de bronzone.

N.B. Het DIM kent voor deze ongemaskeerde persoonsgegevens géén eigen archieffunctie, aangezien het archiveren van operationele gegevens de verantwoordelijkheid is van de bron, en daar dus ook al wordt uitgevoerd.
Daarnaast gelden voor informatiegebieden met ongemaskeerde persoonsgegevens géén langlopende traceerbaarheidseisen, aangezien deze categorie informatiegebieden niet bedoeld is voor gebruik binnen S&V.

Het DLM van de **overige categorieën** informatiegebieden kent een eigen dynamiek, aangezien hier niet alleen het bron-DLM moet worden meegenomen, maar ook het gebruik van de gegevens binnen rapportage/analyse-processen. Ook geldt dat het DIM hier, in tegenstelling tot voor de ongemaskeerde gegevens, wél een archieffunctie heeft (of kan hebben).

N.B. De bewaartermijnen voor klantgerichte processen (en dus ongemaskeerde gegevens) zijn nooit langer dan die voor analyse/rapportage-processen (en dus gemaskeerde gegevens). Vandaar dat voor de verwijdering van anonieme en overige gegevens de bewaartermijnen voor analyse/rapportage kunnen worden aangehouden.

14.4.1 Verwijdering ongemaskeerde persoonsgegevens

Binnen de **integratiezone** loopt het DLM van de **ongemaskeerde persoonsgegevens** gelijk met het DLM van de ongemaskeerde identificerende gegevens (in de **bronzone**) waar de betreffende gegevens (mede) op gebaseerd zijn.[111]

Omdat in de **integratiezone** géén hub/satelliet-structuren worden gebruikt om identificerende en niet-identificerende gegevens van elkaar te scheiden (zoals in de **bronzone**) betekent verwijdering van ongemaskeerde persoonsgegevens hier **ook** verwijdering van de niet-identificerende attributen binnen die persoonsgegevens. Beide typen attributen zitten immers in dezelfde tabel.
Het DLM van de bronzone wordt gevolgd door bij het laden/aanvullen van de gegevens in de integratiezone altijd een delta-proces uit te voeren dat ervoor zorgt dat, voor afleidingen die (mede) zijn gebaseerd op identificerende gegevens in die bronzone, een verwijdering van die data (uit de bronzone) altijd resulteert in een verwijdering van de afleiding (fysiek) in het informatiegebied.
Deze verwijderingen kennen geen aparte goedkeurings- of vastleggingsstap, aangezien het DIM voor operationele gegevens geen formele archief-rol heeft.

N.B. De laadlogica integratiezone voert deze verwijdering in twee stappen uit: eerst worden de gegevens voorzien van een administratieve einddatum/tijd en gemarkeerd als "te vernietigen", en vervolgens worden, als een aparte stap in dezelfde ETL-keten, alleen **ongemaskeerde persoonsgegevens** met een vernietigings-markering fysiek geschoond. Door deze tweetraps-benadering blijft de laadlogica voor gemaskeerde en ongemaskeerde persoonsgegevens, op de laatste stap na (zie volgende paragrafen), identiek.

[111] Als een informatiegebied in de integratiezone (deels) gebaseerd is op andere informatiegebieden, dan loopt de relatie met de bronzone via deze andere informatiegebieden.

N.B. De "grace periode" [112] die in de bronzone wordt aangehouden is niet van toepassing op de integratiezone. Deze haalt zijn gegevens immers via de ontkoppelviews op uit de bronzone, en in die ontkoppelviews worden in de bron verwijderde gegevens direct onzichtbaar (dus niet pas na de grace periode).
Bij onterecht doorgeven van verwijderingen (door de bron) moet de verwijdering in het informatiegebied dus weer ongedaan gemaakt worden. Het eerder genoemde delta-proces draagt hier automatisch zorg voor.[113]

14.4.2 Verwijdering gemaskeerde persoonsgegevens

Het DLM van de **gemaskeerde** persoonsgegevens in het DIM kent een eigen dynamiek, aangezien hier niet alleen het bron-DLM moet worden meegenomen, maar ook het gebruik van de gegevens binnen rapportage/analyse-processen. Ook geldt dat het DIM hier, in tegenstelling tot voor de ongemaskeerde gegevens, wél een archieffunctie heeft. Wel is er een duidelijk verband tussen de bewaartermijnen voor de gemaskeerde gegevens in de bronzone (die immers ook al deels gebaseerd zijn op de behoeften van de rapportage/analyse-processen) en die van de daarvan afgeleide gegevens in de integratiezone.

Verwijdering van gegevens uit de integratiezone volgt dus **niet automatisch** uit de verwijdering van de onderliggende data uit de bronzone, maar is er wel nauw mee verbonden.

Verwijdering van de gemaskeerde gegevens wordt daarom periodiek door een combinatie van stuur-metadata gedreven DIM-processen en handmatige stappen uitgevoerd, als volgt:

1. Bij het laden/aanvullen van gegevens in de integratiezone worden gegevens die (deels) gebaseerd zijn op niet meer in de (ontkoppelviews op de) bronzone beschikbare gegevens zowel voorzien van een administratieve einddatum/tijd als gemarkeerd als "te vernietigen"
2. De DIM-afnemers valideren (bijvoorbeeld m.b.v. BI-tools) dat deze gegevens inderdaad vernietigd kunnen worden.
3. Middels scripts worden alle gegevens waarvoor dit het geval is gemarkeerd met "vernietiging akkoord". Voor gegevens die, om welke reden dan ook, toch nog niet verwijderd mogen worden de vernietigingsmarkering vervangen door "vernietiging uitgesteld".[114]
4. De definitieve vernietigingslijst wordt opgesteld (door FB-DIM)
5. Na goedkeuring vernietiging (door afnemers en bronnen) wordt deze definitieve lijst (of een samenvatting ervan) opgeslagen in het Elektronisch archief (handmatig proces, conform UWV-standaarden op dit gebied)
6. Een automatisch DIM-proces verwijdert de met "vernietiging akkoord" gemarkeerde rijen uit de integratiezone.

Anders dan bij de ongemaskeerde persoonsgegevens (zie vorige paragraaf) worden gegevens dus niet direct na het markeren fysiek geschoond.
Periodiek (vermoedelijk jaarlijks) wordt voor de met "vernietigen uitgesteld" gemarkeerde gegevens, gecontroleerd of deze nog steeds bewaard dienen te blijven. Is dat inderdaad zo, dan blijft de markering staan, is dat niet zo, dan wordt de markering (middels scripts) weer teruggezet

[112] Zie paragraaf 8.3.3 (Volgen van het DLM van de bron: harde en zachte verwijderingen)
[113] Deze "brute force" benadering kan wel een aanzienlijke doorlooptijd vergen. Omdat de grace periode een "vangnet" is voor leveringsfouten die eigenlijk niet mogen voorkomen is dat toch te prefereren boven een slimmere, maar daardoor minder complexere en minder robuuste, heropbouw.
[114] Deze expliciete markering zorgt ervoor dat die gegevens niet bij de eerstvolgende verversing van de integratiezone opnieuw worden gemarkeerd als "te vernietigen"

naar "te vernietigen" zodat de stappen 2-6 hierboven weer kunnen worden doorlopen.

14.4.3 Verwijdering niet-persoonsgegevens

Het proces voor andere dan persoonsgegevens is gelijk aan dat voor gemaskeerde persoonsgegevens.

Een bijzonder geval zijn afgeleide gegevens die wel gebaseerd zijn op persoonsgegevens uit de bronzone, maar binnen de afleiding anoniem zijn geworden, bijvoorbeeld doordat ze over groepen personen heen geaggregeerd zijn.

Strikt genomen zijn deze aggregaten geen persoonsgegeven meer, en zouden ze dus altijd het verwijderproces voor gemaskeerde gegevens moeten doorlopen, ook als ze gebaseerd zijn op ongemaskeerde gegevens uit de bronzone. Dit kan echter een probleem opleveren als de aggregaten consistent moeten blijven met gedetailleerde (niet geaggregeerde) ongemaskeerde gegevens.

Waar deze consistentie-eis inderdaad bestaat wordt deze gedekt door in de informatieproducten in de bedrijfszone alle als "te vernietigen" gegevens weg te filteren.

14.5 DLM en overige archivering/vernietiging in de bedrijfszone

Net als de **integratiezone** bevat de **bedrijfszone** vier categorieën gegevens:

a) ongemaskeerde persoonsgegevens
b) gemaskeerde persoonsgegevens
c) door aggregatie anoniem geworden gegevens
d) overige gegevens (überhaupt geen betrekking hebbend op personen)

Anders dan in de integratiezone bevat de bedrijfszone deze gegevens echter zowel in databases (m.n. datamarts) als in bestanden (bestandsleveringen).

DLM en archivering/vernietiging zijn in de bedrijfszone qua algemene aanpak dus vergelijkbaar met die in de integratiezone, maar verschillen licht per type informatiegebied.

===

14.5.1 Datamarts

De DLM-eisen voor datamarts zijn gelijk aan die voor de integratiezone. De aanpak is dus ook (vrijwel) gelijk.

14.5.1.1 Verwijdering ongemaskeerde persoonsgegevens

Binnen de datamarts loopt het DLM van de **ongemaskeerde** persoonsgegevens gelijk met het DLM van de onderliggende ongemaskeerde identificerende gegevens (in de **bronzone**) en/of ongemaskeerde persoonsgegevens (in de **integratiezone**) waar de betreffende gegevens (mede) op gebaseerd zijn.

Omdat in de **datamarts** géén hub/satelliet-structuren worden gebruikt om identificerende en niet-identificerende gegevens van elkaar te scheiden (zoals in de **bronzone**) betekent verwijdering van ongemaskeerde persoonsgegevens hier **ook** verwijdering van de niet-identificerende attributen binnen die persoonsgegevens. Beide typen attributen zitten immers in dezelfde tabel.

Het DLM van de bronzone wordt gevolgd door bij het laden/aanvullen van de gegevens in de datamart altijd een delta-proces uit te voeren dat ervoor zorgt dat, voor afleidingen die (mede) zijn gebaseerd op identificerende gegevens in die bronzone (direct of via de informatiezone), een

verwijdering van die data (uit de bronzone resp. de informatiezone) altijd resulteert in een verwijdering van de afleiding (fysiek) in de datamart.

Deze verwijderingen kennen geen aparte goedkeurings- of vastleggingsstap, aangezien het DIM voor operationele gegevens geen formele archief-rol heeft.

N.B. De laadlogica voor de datamart voert deze verwijdering in twee stappen uit: eerst worden de gegevens voorzien van een administratieve einddatum/tijd en gemarkeerd als "te vernietigen", en vervolgens worden, als een aparte stap in dezelfde ETL-keten, alleen **ongemaskeerde persoonsgegevens** met een vernietigings-markering fysiek geschoond. Door deze tweetraps-benadering blijft de laadlogica voor gemaskeerde en ongemaskeerde persoonsgegevens, op de laatste stap na (zie volgende paragrafen), identiek.

N.B. De "grace periode" [115] die in de bronzone wordt aangehouden is niet van toepassing op de integratiezone en de bedrijfszone, en dus ook niet op de datamarts.
Bij onterecht doorgeven van verwijderingen (door de bron) moet de verwijdering in de datamart dus weer ongedaan gemaakt worden. Het eerder genoemde delta-proces draagt hier automatisch zorg voor.

14.5.1.2 Verwijdering gemaskeerde persoonsgegevens

Het DLM van de **gemaskeerde** persoonsgegevens in het DIM kent een eigen dynamiek, aangezien hier niet alleen het bron-DLM moet worden meegenomen, maar ook het gebruik van de gegevens binnen rapportage/analyse-processen. Ook geldt dat het DIM hier, in tegenstelling tot voor de ongemaskeerde gegevens, wél een archieffunctie heeft. Wel is er een duidelijk verband tussen de bewaartermijnen voor de gemaskeerde gegevens in de bronzone (die immers ook al deels gebaseerd zijn op de behoeften van de rapportage/analyse-processen) en die van de daarvan afgeleide gegevens in de bedrijfszone.

Verwijdering van gegevens uit de datamart volgt dus **niet automatisch** uit de verwijdering van de onderliggende data uit de bronzone, maar is er wel nauw mee verbonden

Verwijdering van de gemaskeerde gegevens wordt daarom periodiek door een combinatie van stuur-metadata gedreven DIM-processen en handmatige stappen uitgevoerd, als volgt:

1. Bij het laden/aanvullen van gegevens in de datamart worden gegevens die (deels) gebaseerd zijn op niet meer in de (ontkoppelviews op de) bronzone beschikbare gegevens zowel voorzien van een administratieve einddatum/tijd als gemarkeerd als "te vernietigen"
2. De DIM-afnemers valideren (bijvoorbeeld m.b.v. BI-tools) dat deze gegevens inderdaad vernietigd kunnen worden.
3. Middels scripts worden alle gegevens waarvoor dit het geval is gemarkeerd met "vernietiging akkoord". Voor gegevens die, om welke reden dan ook, toch nog niet verwijderd mogen worden de vernietigingsmarkering vervangen door "vernietiging uitgesteld".[116]
4. De definitieve vernietigingslijst wordt opgesteld (door FB-DIM)
5. Na goedkeuring vernietiging (door afnemers en bronnen) wordt deze definitieve lijst (of een samenvatting ervan) opgeslagen in het Elektronisch archief (handmatig proces, conform UWV-standaarden op dit gebied)
6. Een automatisch DIM-proces verwijdert de met "vernietiging akkoord" gemarkeerde rijen uit de datamart.

[115] Zie paragraaf 8.3.3 (Volgen van het DLM van de bron: harde en zachte verwijderingen)

[116] Deze expliciete markering zorgt ervoor dat die gegevens niet bij de eerstvolgende verversing van de integratiezone opnieuw worden gemarkeerd als "te vernietigen"

Anders dan bij de ongemaskeerde persoonsgegevens (zie vorige paragraaf) worden gegevens dus niet direct na het markeren als "te vernietigen" fysiek geschoond.
Periodiek (vermoedelijk jaarlijks) wordt voor de met "vernietigen uitgesteld" gemarkeerde gegevens, gecontroleerd of deze nog steeds bewaard dienen te blijven. Is dat inderdaad zo, dan blijft de markering staan, is dat niet zo, dan wordt de markering (middels scripts) weer teruggezet naar "te vernietigen" zodat de stappen 2-6 hierboven weer kunnen worden doorlopen.

14.5.1.3 Verwijdering niet-persoonsgegevens

Het proces voor andere dan persoonsgegevens is gelijk aan dat voor gemaskeerde persoonsgegevens.

Een bijzonder geval zijn afgeleide gegevens die wel gebaseerd zijn op persoonsgegevens uit de bronzone, maar binnen de afleiding anoniem zijn geworden, bijvoorbeeld doordat ze over groepen personen heen geaggregeerd zijn.

Strikt genomen zijn deze aggregaten geen persoonsgegeven meer, en zouden ze dus altijd het verwijderproces voor gemaskeerde gegevens moeten doorlopen, ook als ze gebaseerd zijn op ongemaskeerde gegevens uit de bronzone. Dit kan echter een probleem opleveren als de aggregaten consistent moeten blijven met gedetailleerde (niet geaggregeerde) ongemaskeerde gegevens.

Waar deze consistentie-eis inderdaad bestaat wordt deze gedekt door bij het laden van de datamart alle als "te vernietigen" gegevens weg te filteren. Als deze eis niet bestaat is herberekening van de dergelijke aggregaten bij verwijdering van de onderliggende ongemaskeerde gegevens niet noodzakelijk.

14.5.2 **Bestandsleveringen**

14.5.2.1 Archivering bestandsleveringen

Het archief bestandsleveringen is alleen bedoeld om herleveringen te kunnen ondersteunen; de afnemer (of de afnemende applicatie) is zelf verantwoordelijk voor de archivering conform Archiefwet.

De bewaartermijn in dit archief is vastgelegd in de SNO en zal, naar verwachting ongeveer gelijk zijn aan de maximale periode tussen twee bestandsleveringen.

Bestanden waarvoor de bewaartermijn is verstreken worden vernietigd.
Dit gebeurt middels een apart, bij voorkeur metadata-gedreven, proces, waarvan de frequentie nog bepaald moet worden.

14.5.2.2 Vernietiging gearchiveerde bestandsleveringen

Bestanden waarvoor de bewaartermijn is verstreken worden vernietigd.
Dit gebeurt middels een apart, bij voorkeur metadata-gedreven, proces, waarvan de frequentie nog bepaald moet worden.

14.5.2.3 DLM op het data-spoor

In **het data-spoor**[117] worden, indien nodig, de gegevens in het bestand (of de gegevens die punt waren bij het aanmaken van het bestand) in gestructureerde vorm opgeslagen.

[117] Vergelijkbaar met de ANL-laag van DWH3.

Deze opslag is bedoeld voor traceerbaarheid en, in sommige gevallen, het kunnen creëren van delta-leveringen.
Ook eventuele invoer-parameters worden, t.b.v. traceerbaarheid en reproduceerbaarheid, in dit data-spoor opgeslagen.

De bewaartermijn voor deze gegevens is vastgelegd in (of af te leiden uit) de SNO, en opgeslagen in de stuur-metadata.
Gegevens waarvoor de bewaartermijn is verstreken worden verwijderd uit het data-spoor.
Dit gebeurt door de DLM-tools bedrijfszone, in een nog te bepalen frequentie.

14.5.3 Gegevensvensters

Archivering en vernietiging is voor gegevensvensters niet relevant; een gegevensvenster is een "virtueel" informatieproduct, en dus is de inhoud volledig afhankelijk van het DLM van de onderliggende onderdelen van het DIM.

14.5.4 Zandbakken

Een zandbak binnen het DIM wordt (op dit moment) geïmplementeerd als een gegevensvenster, met daarin vaak ook informatiegebieden en/of (in 3NF-gemodelleerde) datamarts.

Archivering en vernietiging is voor gegevensvensters niet relevant (zie vorige paragraaf), en dus ook niet voor gegevensvensters t.b.v. zandbakken.

Voor een zandbak buiten het DIM is de afnemer verantwoordelijk voor archivering en vernietiging.
Binnen het DIM zijn hier dus geen processen voor ingericht.
Wel kunnen er in de SNO eisen hieromtrent zijn vastgelegd.

N.B. In de toekomst, als UWV ook de beschikking heeft over technologie voor niet-relationele zandbakken, zal een zandbak wél een apart geïmplementeerd informatieproduct worden.
Dit is ook het geval als Gegevensdiensten fysieke zandbakken (ook relationele) gaat hosten.
Archiverings- en vernietigingsprocessen voor zandbakken kunnen dan wel relevant worden.

14.6 DLM en overige archivering/vernietiging in de end-user zone

De afnemers zijn verantwoordelijk voor alle archivering en vernietiging in de end-user zone.
Binnen het DIM zijn hier dus geen processen voor ingericht.
Wel kunnen er in de SNO eisen hieromtrent zijn vastgelegd.

14.7 Archivering/vernietiging van metadata

Voor de meeste metadata geldt de bewaartermijn voor "Ontwikkelen, onderhouden en uitfaseren van applicaties", dus 5 jaar na uitfaseren van de bron(nen) en/of informatieproduct(en) waar de metadata betrekking op heeft.

Archivering/vernietiging zal hiervoor met handmatige processen worden uitgevoerd.

Log-metadata is alleen van belang voor het DIM zelf (traceerbaarheid, performance-analyse, audit) en zal, vooral om performance-redenen, periodiek geautomatiseerd geschoond worden.

14.8 Vernietiging van backups

Voor opslag en vernietiging van database- en bestands-backups volgt het DIM de UWV–standaards.

15 GENERIEKE OPLOSSINGEN

Het gebruik maken van generieke oplossingen/bouwblokken/harnassen is een "best practice" methode om ervoor te zorgen dat alles consistent wordt uitgevoerd, en dat niet elke keer het wiel opnieuw uitgevonden wordt.

De realisatie hiervan heeft een hogere initiële inspanning, maar deze inspanning wordt later terug verdiend.

Voorwaarde hierbij is wel dat de generieke oplossing zelf weer zo simpel mogelijk gehouden wordt:

- De complexiteit van de generieke oplossing mag niet zo hoog worden dat deze niet meer (of niet meer door anderen dan de oorspronkelijke ontwikkelaar) onderhoudbaar is
- De extra kosten voor de generieke oplossing mogen nooit hoger zijn dan de baten van hergebruik.

Generiek moet dus alleen toegepast worden als het duidelijke toegevoegde waarde biedt.

Een aantal handvatten hiervoor zijn:

- Ontwikkel alleen een generieke oplossing als duidelijk is dat deze vaak (minstens vier keer) gebruikt gaat worden
- Zorg dat de generieke oplossing een "black box" is met duidelijk gedefinieerde (en gedocumenteerde) functionaliteit, en dus ook kan worden toegepast door ontwikkelaars die de interne werking van de generieke oplossing niet kennen
- Zorg dat het aantal parameters om die generieke oplossing te besturen beperkt blijft, en dat hun functie eenvoudig is uit te leggen (weer: ook aan ontwikkelaars die de interne werking van de generieke oplossing niet kennen).
- Probeer niet om alles in één generieke oplossing te vangen; kies liever een 80/20-benadering. Dat kan ook door binnen de generieke oplossing weer een "stopcontact" voor maatwerk in te bouwen (zoals in de harnassen)
- Zorg dat de generieke oplossing functioneel én technisch goed gedocumenteerd is, en valideer dat de oplossing (samen met die documentatie) begrijpelijk én wijzigbaar is voor andere ontwikkelaars.

Binnen het DIM worden de volgende generieke delen onderkend:

- ETL-bouwblokken
- ETL-harnassen
- Maskeringsbouwstenen
- Master sequences

15.1 ETL-bouwblokken

Om zo veel mogelijk tot een standaardisatie van de ETL te komen, zullen er verschillende standaard ETL-bouwblokken worden gedefinieerd die tezamen dan een ETL-proces kunnen vormen.

Idealiter zijn ETL-bouwblokken zo ingericht dat de DataStage-ontwikkeltools ze "drag and drop" kunnen opnemen in ETL-jobs.

Dit document zal geen detailbeschrijving van deze ETL-bouwblokken bevatten. Deze zullen in de technische documentatie behorende bij de ETL-harnassen en bouwstenen worden vastgelegd. Bij de bouwstenen moet je denken aan specifieke stukken functionaliteit die hergebruikt kan worden. Voorbeelden:

- Controle of een bepaald bestand bestaat
- Loggen van het starten van een processtap binnen een ETL-harnas.

ETL-bouwblokken worden ook gebruikt voor zaken als proces-logs, fout-logs, communicatie vanuit een harnas naar DIM beheer, archivering van bronleveringen, bepaling omgevingsparameters, refreshen van ontkoppelviews.

De lijst ETL-bouwblokken is continu aan verandering onderhevig, en wordt dus niet in dit ontwerp opgenomen. Ze is wel zichtbaar in de ontwikkelomgeving.

15.2 ETL-harnassen

Wat wij een ETL-harnas noemen is een gestandaardiseerde op stuur-metadata gebaseerde ETL-procesflow bestaande uit bouwblokken die gegevens uit een laag naar de volgende laag transporteert en eventueel transformeert.
Uitgangspunt is om zo veel mogelijk van de ETL-procesflows gestandaardiseerd te hebben.

Voordelen zijn:

- Reductie van de complexiteit doordat alles een gelijke opbouw heeft
- Metadata changes in plaats van source code changes (indien mogelijk)
- Betere onderhoudbaarheid
- Kortere doorlooptijd om nieuwe data in de bronzone beschikbaar te maken (alleen metadata deployment)

De bedoeling is dat indien mogelijk de ETL-harnassen parallel processing van data faciliteren.

Meer functioneel detail over de diverse ETL-harnassen in de hoofstukken over de zones waarin ze gebruikt worden.

15.3 Maskeringsbouwstenen

Om te zorgen dat de maskering van een specifiek functioneel veld ook altijd op dezelfde manier uitgevoerd wordt maken we gebruik van een specifieke vorm ETL-bouwblok: de maskeringsbouwsteen.

In de stuur-metadata wordt vervolgens vastgelegd welke maskeringsbouwsteen op welk veld moet worden toegepast.

Een maskeringsbouwsteen is, voor het aanroepende proces, een "black box", met één input (de te maskeren waarde), en één output (het resultaat van de maskering).

Meestal is een maskeringsbouwsteen een "wrapper" met daarbinnen een call van een Optim-functie, met eventueel wat voor- en nawerk.

De maskeringsbouwstenen kunnen direct op attributen worden toegepast (dus aangeroepen door het bronzone-harnas) of indirect, als onderdeel van row level masking.

Meer details over de diverse maskeringsbouwstenen in de technische documentatie en in het spreadsheet **Bescherming van persoonsgegevens in het DIM - Matrix Maskeringsklassen**.

15.4 Master sequences

Een master sequence is een geparametriseerd ETL-raamwerk (gebouwd m.b.v. DataStage) dat de losse modules binnen een ETL-keten (harnassen, ETL-jobs, soms scripts) in de juiste volgorde aanroept.

Hierdoor wordt het geheel eenvoudig in te plannen, en blijft de in IWS opgenomen intelligentie zo beperkt mogelijk, zoals beschreven in paragraaf 13.1 (Scheduling).

De twee op dit moment onderkende master sequences zijn:

- De master sequence bronverwerking (zie paragraaf 8.6)
- De master sequence informatievoorziening (zie paragraaf 9.4)

In de toekomst worden er additionele master sequences verwacht. Ook is het mogelijk dat de master sequence informatievoorziening wordt gesplitst in aparte master sequences, bijvoorbeeld per type informatieproduct, of apart voor informatiegebieden (integratiezone) en informatieproducten (bedrijfszone).

16 ONDERSTEUNING SELF SERVICE BI & ANALYSE

De eerste onderzoeken naar de functionele, technische en procedurele vereisten voor Self Service lopen nog[GL(27]:

- PoV GINA / Uitkeren
- Inrichting BO repositories (w.o. ook de inrichting van de door BusinessObjects-universes gebruikte systeemaccounts en de eisen aan self service universes)

Daarnaast lijkt PowerBI de voornaamste BI-tool voor self service te worden. UWV-brede inrichting van die tool valt buiten het project DataFabriek, en er is nog weinig over bekend.

Ondersteuning van self service bij het WERKbedrijf, tenslotte, is onderdeel van de migratie van DWH2, en valt daarmee na fase 5.

Dit hoofdstuk zal deels in de loop van fase 4 verder worden ingevuld, en deels in fase 5.

17 TECHNISCHE INFRASTRUCTUUR

17.1 Tiers

In het rekencentrum zijn, voor het DIM, de volgende "tiers" ingericht:

- Client Tier
- Services Tier
- Engine Tier
- Database Tier
- Metadata bridge Tier
- Enterprise search Tier

(detailinformatie hierover in de HLD)

M.u.v. (nu nog) de Database Tier is deze infrastructuur volledig gescheiden voor de vier OTAP-omgevingen.

De topologie van deze vier omgevingen is gelijk, de sizing verschilt.

Wel is de sizing van A en P gelijk, zodat A:

- bij rekencentrum-calamiteiten, eventueel kan worden gebruikt als fail-over voor P
- bij acceptatietests een realistisch beeld geeft van de in P te verwachten performance en gegevenscomplexiteit

N.B. Vergroten van de sizing van m.n. de Engine en Database Tiers kan significante impact hebben op de licentiekosten (bij IBM respectievelijk Oracle).

17.1.1.1 Client Tier

De Client Tier is ingericht in de bestaande UCRA-omgevingen.

Zie paragraaf 17.2.2 (Client-tools - UCRA en KA) voor meer details.

17.1.1.2 Services Tier

Hierop draaien de algemene services van IBM InfoSphere Information Server (de "ETL+"-tooling).

Deze services draaien binnen IBM WebSphere Application Server (WAS).

N.B. IBM WAS is integraal onderdeel van de InfoSphere-licentie. Hieraan zijn dus **geen** aparte licentiekosten verbonden (ook niet als interne doorbelasting)

17.1.1.3 Engine Tier

Het hart van de ETL-verwerking (voorzover die niet "pushed down" is naar de database).

De engine tier acteert ook als de "eigenaar" van de bestandssystemen van het DIM (anders dan bij de legacy DWH's, waar de database tier die rol heeft).

In de HLD wordt dat beschreven als de "Shared Storage". Deze shared storage is ook benaderbaar vanaf de Database Tier.

De belangrijkste onderdelen van de shared storage zijn:

- De folders voor het verwerken van bronleveringen (inbox, verwerkingsfolder, hot archive bronleveringen)
- Het cold archive voor bronleveringen (in een apart file system)
- De folders voor het verspreiden van bestanden geleverd door "bestandslevering"-informatieproducten (server- en transit-outbox, archief bestandsleveringen)

N.B. De transit-outboxen dienen, per outbox geautoriseerd, ook benaderbaar te zijn vanuit de KA-omgeving (al dan niet via UCRA). Potentieel geldt iets vergelijkbaars voor inboxen, bijvoorbeeld als die gebruikt worden voor een "bron van verbeteringen"[118].

17.1.1.4 Database Tier

Het DIM gebruikt voor Oracle de bestaande AIX-infrastructuur voor DWH2/3.

Bij de migratie naar DXC zal deze infrastructuur verhuizen naar Linux, en meteen ook gesplitst worden in aparte infra voor alle vier de OTAP-omgevingen.[119]

17.1.1.5 Metadata bridge Tier

Deze Tier wordt gebruikt bij het importeren van metadata uit tools, bestanden, en databases in de metadata repository van InfoSphere Information Server.

N.B. De enige Windows-server. Alle andere servers draaien Linux (of gaan dat bij DXC draaien).

17.1.1.6 Enterprise search Tier

Deze tier wordt gebruikt bij metadata management.

17.2 Gebruikte (standaard)software

17.2.1 Server-side

Deze paragraaf wordt in de eerstvolgende [GL(28]*versie van dit ontwerp toegevoegd, en dekt zowel de ETL-software als, biivoorbeeld, IWS.*

Gezamenlijke actie voor applicatiearchitect, DBA's en principal ontwikkelaars.

17.2.2 Client-tools - UCRA en KA

Deze paragraaf wordt in de eerstvolgende versie van dit ontwerp toegevoegd, en dekt zowel ontwikkel- en beheer-software als gebruik/inrichting van transit-folders.

Gezamenlijke actie voor applicatiearchitect, DBA's en principal ontwikkelaars.

[118] Zie Bijlage G: Datakwaliteitsbeheer, onder Overschrijven kan niet, verbeteren wel

[119] Op dit moment draaien de O- en T-omgevingen voor Oracle beiden op de T-infra

18 INFORMATIEBEVEILIGING EN –BEHEER

18.1 Strikter ingericht

Vergeleken met de bestaande situatie (bij DWH1/2/3) zijn informatiebeveiliging en informatiebeheer strikter ingericht.

Voorbeelden daarvan zijn:

- Toegangsbeheer en beveiliging "dichter op de database"
- Toegang tot A/P (voor afnemers én afnemende applicaties) alleen via de bedrijfszone, en alleen als rechtsgrond/proportionaliteit/subsidiariteit is aangetoond
- Toegang tot A/P (voor ontwikkelaars/beheerders) alleen bij calamiteiten, en altijd middels een "red envelope"-procedure
- Toegang tot A/P (voor DBA's en, eventueel, release management door andere DWH-medewerkers) strikt gereguleerd (en gemonitord)
- Gebruik van A/P (door afnemers én afnemende applicaties) beter/preciezer gelogd en gemonitord

Deze voorbeelden worden hieronder verder uitgewerkt.

Voor de overige aspecten m.b.t. informatiebeveiliging en informatiebeheer wordt verwezen naar de bestaande praktijk voor DWH1/2/3.

18.1.1 Toegangsbeheer en beveiliging "dichter op de database"

In de bestaande situatie ligt vrij veel toegangsbeheer en –beveiliging binnen de BI-tools (m.n. BusinessObjects).

Waar technisch mogelijk zal dit worden verschoven naar (of ook worden uitgevoerd op) het Oracle-niveau, om de volgende redenen:

- Strikter, en eenvoudiger te monitoren, toegangsbeheer
- Betere ondersteuning van self service, en daardoor een grotere rol voor gegevensvensters, al dan niet als basis voor in self service gebouwde BusinessObjects-universes
- Breder gebruik van PowerBI
- Beter inzicht in het gegevensgebruik van de BI-users zelf (en niet alleen van de verzamelde users van een door die BI-tool geboden product)

Vanwege bovenstaande neemt het aantal Oracle-accounts, en daarmee de voor toegangsbeheer vereiste inspanning, waarschijnlijk sterk toe. In fase 5 van het project zal hiervoor. o.b.v. de met de eerste informatieproducten geleerde lessen, tooling worden ingericht.

18.1.1.1 BusinessObjects

Voor BusinessObjects loopt de komende [GL(29)*maanden een onderzoek naar de (on)mogelijkheden van verbeterd toegangsbeheer. Onderstaande dus onder voorbehoud.*

Voor BusinessObjects zullen waarschijnlijk de volgende verbeteringen worden doorgevoerd:

- Van één systeemaccount naar een systeemaccount per samenhangende groep informatiegebieden en/of datamarts.
  Hoogstwaarschijnlijk door op die groep een gegevensvenster te definiëren, en het systeemaccount daar toegang toe te geven.
- Gebruikersgegevens door BusinessObjects meegegeven met query.
  Deze informatie zal m.n. gebruikt worden voor logging/monitoring.
  Potentieel kan e.e.a. ook gebruikt worden voor toegangsvalidatie en/of VIP/BZ-filtering.

- Eenduidiger afspraken m.b.t de eisen waaraan in self service gebouwde BusinessObjects-universes moeten voldoen, en borging daarvan middels de SNO

18.1.1.2 PowerBI

Het is nog onduidelijk hoe de, door het DIM te ondersteunen, UWV-brede inrichting van PowerBI eruit gaat zien. Het inrichtingsproject daarvoor (buiten het project DataFabriek) is namelijk nog in een pril stadium. De afdeling DWH zal, mede namens het project, de eisen aan de PowerBI-inrichting neerleggen bij het project. Deze eisen zullen in grote lijnen dezelfde zijn als die voor de BusinessObjects-inrichting.

18.1.1.3 Self service

Bij Self service zijn de voornaamste aandachtspunten procedureel.
De belangrijkste daarvan zijn:

- de mate waarin de self service partij zelf het toegangsbeheer mag/moet uitvoeren.
- de garantie dat de self service partij compliant blijft met de voor het DIM én haar afnemers vastgelegde processen en procedures

Inrichting van processen en procedures wordt door de lijn uitgevoerd, en valt buiten de project-scope.

18.1.2 Toegang alleen via de bedrijfszone, en alleen bij "rechtsgrond"

Toegang tot de Acceptatie- en Productie-omgevingen van het DIM loopt, voor afnemers én afnemende applicaties, altijd via de bedrijfszone.

Toegang wordt alleen verleend als er een SNO is. In die SNO is ook vastgelegd dat het afnemende proces (of de afnemende applicatie) rechtsgrond heeft voor het gebruik van de ter beschikking gestelde data, en dat dat gebruik voldoet aan de AVG-principes m.b.t. proportionaliteit en subsidiariteit, en niet in tegenspraak is met eventuele door de bron aangegeven gebruiksbeperkingen.

Opstelling van een nieuw SNO-sjabloon, en inrichting van de daaraan gerelateerde processen wordt door de lijn uitgevoerd, en valt buiten de project-scope.

18.1.3 IV-toegang alleen bij calamiteiten

Toegang tot de Acceptatie- en Productie-omgevingen van het DIM is, voor ontwikkelaars en beheerders, alleen toegestaan voor het oplossen van calamiteiten.
Deze toegang is tijdelijk, en loopt via een "red envelope"-procedure.

Beheerders hebben uiteraard wel toegang tot de monitoring-dashboards in Acceptatie- en Productie.

Inrichting van de "red envelope"-procedure wordt door de lijn uitgevoerd, en valt buiten de project-scope.

Bovenstaande geldt ook voor toegang tot de transit-folders. Het verhuizen van bestanden uit een transit-folder naar de KA-omgeving moet dus een geautomatiseerd proces zijn, en geen handmatige (door een medewerker van DWH-Exploitatie uitgevoerde) stap.

Inrichting van dit geautomatiseerd proces wordt onderdeel van het eerste informatieproduct dat een verhuizing naar de KA-omgeving vereist.

18.1.4 Toegang voor DBA's strikt gereguleerd

DBA's hebben voor hun werk toegang tot de Acceptatie- en Productie-omgevingen nodig.
Ook het promoveren van releases naar Acceptatie- en Productie vereist toegang tot die omgevingen.

Deze toegang zal strikt gereguleerd (en gemonitord) worden.

Inrichting van de hieraan gerelateerde processen en procedures wordt door de lijn uitgevoerd, en valt buiten de project-scope.

Het promoveren van releases gebeurt, indien mogelijk, geautomatiseerd.

De mogelijkheden daartoe worden later in fase 5 door het project onderzocht.

18.1.5 Gebruik beter/preciezer gelogd en gemonitord

Het gebruik van de data in de Acceptatie- en Productie-omgevingen (door afnemers én afnemende applicaties) zal beter en preciezer worden gelogd en gemonitord.

De basis van deze logging/monitoring blijft conform de UWV-standaard (verwerking van logs door QRadar).

Daarnaast wordt een aantal extra controles uitgevoerd, met name bedoeld om, in ieder geval technisch, "in control" te zijn m.b.t. AVG-compliance:

- **Welke gebruiker heeft op welk moment welke views geraadpleegd.**
  De uitgevoerde query wordt opgeslagen, en niet het, potentieel privacy-gevoelige, resultaat.
  De controle wordt op database niveau uitgevoerd, bij voorkeur op een voor alle DWH's (ook de legacy) gelijke wijze.
  De controle wordt door de lijn ingericht, en valt buiten scope van het project.

- **Welke autorisaties voldoen niet aan de DIM-regels.**
  Denk hierbij, bijvoorbeeld, aan toegang tot zowel gemaskeerde als ongemaskeerde gegevens, of aan directe toegang tot de bronzone (dus niet via de ontkoppelviews).

  Het eerste zou alleen mogen voorkomen bij de voor de ETL-processen gebruikte systeemaccounts, het tweede zelfs daar maar beperkt.
  Indien er non-compliance wordt vastgesteld zal er melding gemaakt worden bij DIM-beheer.
  Eventuele vervolgacties vallen buiten scope van het project, en dus ook van dit document.

Voor bovenstaande controles is het cruciaal dat BI-tools niet puur met een systeemaccount de database benaderen, maar daarnaast ook gebruikersgegevens doorgeven.

18.2 Secure Software Development (SSD)

Voor SSD-gerelateerde zaken wordt verwezen naar de DIM-ontwikkel-standaarden.

18.3 Business Continuity Management (BCM)

Het Business Continuity Management (BCM) van het DIM dient tenminste de volgende aspecten te dekken:

- Disaster Recovery Plans (DRP's) voor gegevens-reparatie (door roll backs, herdraaien, etc) bij fouten in de bronlevering en/of de ETL-processen die die de integriteit van de DIM-data hebben geraakt.
- DRP's voor het inlopen van achterstallige bronverwerkingen bij bronleveringen die langdurig "uit de lucht" zijn
- DRP's voor infrastructurele calamiteiten bij het Rekencentrum, of in de standaard-software

BCM van de m.b.v. het DIM geleverde producten/diensten valt buiten de project-scope, maar maakt wel gebruik van binnen het project (en deels speciaal daarvoor) ingerichte DIM-functionaliteit.

18.3.1 Hot en cold archief

Beschreven in paragraaf 14.1 (Archivering van bronleveringen~~Archivering van bronleveringen~~)

Splitsing van de archieven in hot en cold storage kan Disaster Recovery versnellen; vaak voldoet de inhoud van het hot archive (naast de database-backups en -logfiles) namelijk voor het weer consistent maken van de bronzone na een infra-calamiteit.

N.B. Op dit moment zijn de hot en cold archieven technisch identiek ingericht. Bij de migratie naar DXC moet gekeken worden of dat daar ook zo blijft en, zo ja, of splitsing van de archieven in hot en cold nog relevant blijft.

18.3.2 Zachte verwijderingen

Beschreven in paragraaf 8.3.3 (Volgen van het DLM van de bron: harde en zachte verwijderingen~~Volgen van het DLM van de bron: harde en zachte verwijderingen~~)

Toepassing van zachte verwijderingen maakt de verwerking van bronleveringen in de bronzone robuuster, maar kan (als die zachte verwijderingen ongedaan moeten worden gemaakt) grote impact hebben op de verwerkingstijden binnen de integratiezone en de bedrijfszone.

Ongedaan maken is noodzakelijk als de zachte verwijdering veroorzaakt bleek door een leverfout van de bron. Het verdient daarom aanbeveling om die kans zo klein mogelijk te maken, allereerst door goede contract-afspraken met de bron (en rigide testen van de bronlevering), maar daarnaast ook door extra controles tijdens de bronverwerking (zie paragraaf 7.4 (Laadlogica staging)).

Monitoring van (het ongedaan maken van) zachte verwijderingen (door DIM-beheer) kan handig zijn om afnemers op tijd te waarschuwen voor eventuele lever-vertragingen.

18.3.3 Functionele backups

Van de DIM-gegevens zal, conform de (voor de afdeling DWH verbijzonderde) standaards van UWV, periodiek een backup worden gemaakt.

Daarnaast is het mogelijk om, middels gerichte deel-backups, de toestand van de DIM-gegevens net vóór of net na hun verversing veilig te stellen, om het ongedaan maken van die verversing te vereenvoudigen.

Het DIM volgt hier de bestaande praktijk binnen de afdeling DWH,

18.3.4 Database-partitionering

De doorlooptijd van gegevens-reparaties (bijvoorbeeld door een restore van een eerdere versie van die gegevens) kan aanzienlijk versneld worden door database-partionering.

Inrichting van die partionering is onderdeel van het detail-ontwerp van de DIM-databases, en valt buiten scope van dit document.

18.3.5 Lever-spoor

Beschreven in paragraaf 10.2.5 (Traceerbaarheid via het lever-spoor~~Traceerbaarheid via het lever-spoor~~)

M.b.v. de gegevens in het **data-spoor** kunnen parameter-gedreven informatieproducten na het repareren van een gegevens-issue opnieuw worden aangemaakt.

Ongewijzigde herlevering van bestanden (bij transport- of afnemersproblemen) wordt ondersteund door het **archief bestandsleveringen**.

Bijlage A: BELANGRIJKE AFKORTINGEN EN BEGRIPPEN

Lijst van afkortingen en begrippen (afkortingen van applicaties zijn te vinden in Confipedia)

Afkorting/term	Verklaring, eventueel met toelichting
AVG	Algemene Verordening Gegevensbescherming. De Nederlandse versie van de GDPR (General Data Privacy Regulation)
BCM	Business Continuity Management (BCM) houdt zich bezig met het inzichtelijk maken van factoren die de bedrijfsprocessen kunnen verstoren en het bepalen van maatregelen om de kans hierop te verkleinen of de gevolgen te beperken.
Bestandslevering	Levering van in meer of mindere mate bewerkte gegevens, doorgaans in de vorm van platte bestanden. Binnen dit ontwerp verwijst "bestandslevering" altijd naar een informatieproduct (dus een levering van het DIM aan een afnemer), en niet naar een levering van een bron aan het DIM. Zie daarvoor "bronlevering". N.B. In de PSA heet dit type informatieproduct "gegevenslevering". Omdat die term binnen DWH en Gegevensdiensten ook voor andere concepten wordt gebruikt, wordt in dit Conceptueel Ontwerp steeds de term "bestandslevering" gebruikt.
BGB	Bijzondere GevalsBehandeling
CGM	Canoniek Gegevens Model
DA GEIN	DoelArchitectuur GEgevensINtegratie. Context voor deze PSA
Datamart	Een op maat gemaakte gegevensverzameling, vaak dimensioneel gestructureerd, ten behoeve van een specifieke gebruikersgroep en/of -proces (meestal rapportages, OLAP, dashboards, of vergelijkbaar), inclusief self service BI. Vaak vereist het op maat maken complexe transformaties en afleidingen
DBA	DataBase Administrator
DF	DataFabriek [GL(30)](het project) of Datafabriek (de organisatie)
DIA	Data Integratie en Analyse
DIM	Data Integratie Magazijn
distro	Een door een leverancier "enterprise ready" gemaakte samenhangende set van open source software componenten, meestal geleverd inclusief support.
DLM	Data Lifecycle Management
DRP	In de context van het DIM is een disaster recovery plan (DRP) een gedocumenteerde en gestructureerde beschrijving van de stappen die moeten worden doorlopen om, na een ongepland incident, zo snel mogelijk weer (met betrouwbare data!) in productie te zijn. DRP's zijn onderdeel van het Business Continuity Management (BCM).
DTL	DaTa-Laag Onderdeel van DWH 3.0, vergelijkbaar met de bronzone van het DIM.
DWARFS	Officiële naam van DWH 1.0
DWH	Data Warehouse De afkorting wordt ook gebruikt voor de beheer-afdeling van de huidige centrale data warehouses, en voor het toekomstige IV-domein waar dit beheer zal gaan landen.
EA GH	Enterprise Architectuur GegevensHuishouding. Document in wording. Dit document zal een groot aantal "open punten" m.b.t. de precieze rol van het DIM, en ander applicaties in het DIA-domein, adresseren.

EP	Eigen Personeel
ELT-patroon	Een manier van gebruik van ETL-software waarbij alle gegevens eerst vanuit een gegevensbron naar het gegevensdoel (meestal een relationele database) worden getransporteerd (Extract, Load), waarna de rekenkracht van die doel-omgeving wordt gebruikt voor de gegevenstransformaties (Transform). Bij het ELT-patroon voert de ETL-software dus niet zelf de transformaties uit (zoals bij het ETL-patroon). In plaats daarvan genereert de ETL-software code (bv. PL/SQL) die binnen de doelomgeving kan worden uitgevoerd.
ETL	Extraction Transformation Loading Als pakket in de markt beschikbare functionaliteit voor gegevenstransformatie en –replicatie, i.h.a. geoptimaliseerd voor gebruik in bulkprocessen.
ETL-patroon	Een manier van gebruik van ETL-software waarbij alle gegevens eerst vanuit een gegevensbron naar de ETL-omgeving worden getransporteerd (Extract), daar worden getransformeerd (Transform) en vervolgens geladen in het gegevensdoel (meestal een relationele database) (Load).
ETL+	ETL-software plus daaraan nauw gerelateerde, ook door UWV-benodigde software t.b.v. gegevensmaskering, datakwaliteitsmanagement en metadata management. *(UWV-specifieke term, gebruikt in de ETL-aanbesteding)*
FO	Functioneel Ontwerp
FUGEM	FUnctioneel GEgevens Model
GAT	GebruikersAcceptatieTest
GEB	GegevensbeschermingsEffectBeoordeling (Engels: Privacy Impact Assessment; PIA)
Gegevensvenster	Een virtuele gegevensverzameling (deelverzameling van de in het DIM beschikbare gegevens) die ontsloten wordt ten behoeve van self-service BI en/of self service analytics, d.w.z. rechtstreeks gebruik door afnemers, bijvoorbeeld t.b.v. querying en ad-hoc rapportage. Anders dan bij datamarts is er geen sprake van complexe transformaties en afleidingen. Wel bevatten gegevensvensters filters op tabel-, rij- en attribuutniveau. Dit om te garanderen dat de getoonde deelverzameling alleen gegevens bevat waarvoor rechtsgrond/proportionaliteit/subsidiariteit geldt.
GLO	GegevensLeveringsOvereenkomst. Een “contract” waarin de afspraken rondom een levering (door een bron, aan het DIM) zijn vastgelegd. N.B. Het GLO-proces wordt, ten tijde van het schrijven van dit ontwerp, aangepast door het Implementatieteam GD/DWH (lijninitiatief, geen onderdeel van het project DataFabriek). Onderdeel van die aanpassing is vervanging van de GLO door een nieuwe contractvorm, met een nieuwe naam. Lees, in dit document, “GLO” dus als “GLO of de opvolger daarvan”.
HRC	HoofdRekenCentrum. Kan zowel verwijzen naar het bestaande rekencentrum (bij IBM) als naar het nieuwe (bij DXC)
HUDSV	Handboek Uniforme Definities Sturen & Verantwoorden
IDA	IBM Infosphere Data Architect
IGC	IBM Infosphere Information Governance Catalog
Informatiegebied	Een deel-model binnen de integratiezone. Elk informatiegebied is gebaseerd op gegevens uit de bronzone (van één of meer bronnen), al dan niet gecombineerd met gegevens uit andere

	informatiegebieden in de integratiezone, en meestal bedoeld voor een specifieke groep informatieproducten. N.B. In de migratiestrategie en de deliverables van het project DataFabriek is ook sprake van informatiegebieden. De term wordt daar gebruikt voor, vanwege hun grote samenhang, in één keer te migreren groepen informatieproducten, al dan niet (deels) gebaseerd op gegevens de integratiezone. Gepoogd zal worden om in de migratiestrategie en deliverables hiervoor de term "migratieblok" te introduceren, om misverstanden in de toekomst zo te beperken.
Informatieproduct	Een door het DIM aan afnemers ter beschikking gestelde dataset. Er zijn vier typen informatieproducten: Datamarts, Gegevensvensters, Bestandsleveringen en Zandbakken.
LDM	Logical Data Model / Logisch Data Model
MVL	Materialized View Laag Onderdeel van DWH 3.0 Een op de DaTa-laag (DTL) van datzelfde DWH gebaseerde set Oracle Materialized Views (alleen die objecten die in de integratiezone nodig zijn). Ligt op het snijvlak van bronzone (DTL) en integratiezone. De inhoud van de MVL is op wat schoning na gelijk aan de DTL, maar wordt in plaatjes vaak onderin de integratiezone getekend.
NGP	Nieuw GegevensPakhuis – Officiële naam van DWH 2.0
PAM	PAM: Privileged Accountmanagement; DXC-oplossing voor het monitoren van gebruikers met risicovolle rechten op applicaties en/of infrastructuur
PDM	Physical Data Model (fysiek datamodel)
Polisdomein	De door Gegevensdiensten beheerde applicaties met (kopieën) van basisadministraties, waaronder POLIS (Loonaangifteketen) en UPA (UWV-kopie/uitbreiding van de BRP)
PUIK	Uniek ID voor elke UWV-medewerker (intern of extern). De afkorting zelf is een overblijfsel uit het verleden (Project Uniformering Inrichting Kantoorautomatisering)
RDBMS	Relationeel DataBase Mangement Systeem
RLO	Record LayOut; een gedetailleerde beschrijving van de door een bron geleverde gegevens.
S&V	Sturen & Verantwoorden
SNO	ServiceNiveauOvereenkomst. Een "contract" waarin de afspraken rondom een levering (door het DIM, aan een afnemer) zijn vastgelegd. N.B. Het SNO-proces wordt, ten tijde van het schrijven van dit ontwerp, aangepast door het Implementatieteam GD/DWH (lijninitiatief, geen onderdeel van het project DataFabriek). Onderdeel van die aanpassing is vervanging van de SNO door een nieuwe contractvorm, met een nieuwe naam. Lees, in dit document, "SNO" dus als "SNO of de opvolger daarvan".
S&V	Sturen en Verantwoorden; UWV-term voor de (de voortbrengingsprocessen van) Management Informatie
TEGEM	TEchnisch GEgevens Model
TCO	Total Cost of Ownership
UDS	UWV Data Store – Officiële naam van DWH 3.0
UGL	Universele GegevensLaag. De laag in DWH 3.0 waar de halffabrikaten in staan. Vergelijkbaar met de integratiezone van het DIM
VK3	Vertrouwelijkheidsklasse 3 – Zeer gevoelige gegevens

Zandbak	Een data-omgeving t.b.v. analyse. Een zandbak kan virtueel of fysiek zijn, of een combinatie van beide.

Bijlage B: RATIONALE BREDE BRONONTSLUITING

Bij het definiëren van de gegevens-scope van een bronontsluiting (de "breedte") zijn drie benaderingen mogelijk:

	Scope	**Aanpassingen als**
Smal	De bronontsluiting bevat alleen de brongegevens waaraan een **expliciete behoefte** is (bij de gegevensafnemers van het DIM).	De bronontsluiting wordt aangepast als nieuwe informatiebehoeften additionele brongegevens vereisen.
Breed	De bronontsluiting bevat alleen de brongegevens waaraan een **verwachte behoefte** is (bij de gegevens afnemers van het DIM).	De bronontsluiting wordt aangepast als nieuwe informatiebehoeften onverwacht additionele brongegevens vereisen, óf als wijzigingen in de bron resulteren in extra gegevens met een verwachte behoefte.
Volledig	De bronontsluiting bevat de **volledige inhoud** van het bronsysteem.	De bronontsluiting wordt bij elke wijziging van de bron aangepast.

Deze drie benaderingen hebben, vanuit beheer-perspectief, de volgende voor- en nadelen:

	Voordelen	**Nadelen**
Smal	➢ Minimale dubbele opslag ➢ Geen impact op DIM als bron wijzigt voor (nog) niet door afnemers gebruikte gegevens ➢ Expliciete doelbinding/rechtsgrond van de replicatie naar (en de opslag in) het DIM	➢ Lange time-to-market voor informatieproducten die nieuwe brongegevens vereisen ➢ Met terugwerkende kracht laden van de historie van nieuw vereiste gegevens meestal niet mogelijk ➢ Veel wijzigingsverzoeken bij bronnen, vaak met hoge urgentie
Breed	➢ Alleen lange time-to-market voor informatieproducten die **onverwacht** nieuwe brongegevens vereisen. ➢ Historie van nieuw vereiste gegevens vaak al beschikbaar	➢ Brede doelbinding/rechtsgrond voor de replicatie naar (en de opslag in) het DIM noodzakelijk.
Volledig	➢ Brongegevens reeds beschikbaar voor alle nieuwe informatieproducten	➢ Volledige dubbele opslag ➢ Doelbinding/rechtsgrond replicatie complex ➢ Grote investering vooraf, zowel bij bron als bij DIM ➢ Elke bronwijziging heeft impact op het DIM.

Vanuit project-perspectief gelden daarnaast de volgende voor- en nadelen:

	Voordelen	**Nadelen**
Smal	Alleen migratie van historische gegevens naar het DIM als deze gegevens nu al in een informatieproduct worden gebruikt.	Scope bestaande bronleveringen (naar de legacy DWH's) moet herijkt worden; niet in een informatieproduct gebruikte gegevens moeten uit de interface worden verwijderd.
Breed	Geen herijking van bronleveringen; de scope van de bestaande leveringen blijft in grote lijnen ongewijzigd. Waar verbreding van een levering toch noodzakelijk blijkt is de impact hiervan beperkt; kan relatief eenvoudig worden meegenomen met de overige vereiste wijzigingen aan die levering.	Vereist aanpassingen aan een deel van de bestaande interfaces.
Volledig	geen	Vereist grote aanpassingen aan alle bronleveringen. Vertraagd daarmee de voortgang van het project, zonder dat hier duidelijke bedrijfswaarde tegenover staat.

Voor het DIM is daarom, zowel binnen het project als voor de langere termijn, gekozen voor brede bronontsluiting:

- Brede bronontsluiting is een "market best practice" die in DWH-omgevingen (zowel binnen als buiten UWV) algemeen wordt toegepast;
- Impact ervan op het DataFabriek-project (qua doorlooptijd en risico) is beperkt, aangezien de bronontsluitingen toch al moeten worden aangepast t.b.v. robuustheid/compliance.
- Om deze impact nog verder te beperken is in de interface-standaarden gekozen voor de benadering "Gegevensobjecten o.b.v. behoefte, gegevensattributen o.b.v. toekomstige behoefte".
  Daarmee wordt bedoeld dat **binnen het project** brede ontsluiting alleen geldt voor nu reeds (aan de "legacy" DWH's) geleverde gegevensobjecten; nu nog niet ontsloten gegevensobjecten worden dus pas opgenomen in een bron-interface als een nieuw DIM-informatieproduct deze, buiten het project om, vereist.
- In het algemeen zal het ontwikkelen van een dergelijk nieuw DIM-informatieproduct, en dus ook de eventuele daarvoor noodzakelijke verbreding van een bronontsluiting buiten scope vallen van het project DataFabriek, en ook niet randvoorwaardelijk zijn voor de voortgang van dat project. Het zal echter wel "business as usual" zijn voor de nieuwe lijnorganisatie Datafabriek.

Bijlage C: RATIONALE MAATWERK STUUR-METADATA

Rationale maatwerk-structuur voor stuur-metadata

De stuur-metadata staat in een als maatwerk ontwikkelde Oracle database, en is dus géén onderdeel van de metadata binnen de diverse metadata-tools van de IBM Infosphere suite.

Gebruik van die tools voor de stuur-metadata is wel onderzocht, maar bleek geen haalbare oplossing:

- Het metadata-model van IBM Infosphere is proprietary. Er zijn geen garanties dat patches en upgrades niet tot wijzigingen in het metadatamodel leiden. Dit zou kunnen leiden tot onverwachte fouten in de ETL-harnassen, waardoor er voor elke patch of upgrade een complete regressietest moet worden uitgevoerd.
- Het metadata-model van IBM Infosphere bevat niet alle voor de ETL-harnassen benodigde informatie. Hiervoor zou er een uitbreiding moeten worden gemaakt welke ergens separaat zou moeten worden opgeslagen. Dit zou het voor de ETL-harnassen complexer maken om de voor hun werking benodigde metadata bij elkaar te brengen.
- De scope van de stuur-metadata verschilt van die van IGC. Enerzijds bevat IGC (of eigenlijk de FUGEM's die er de voornaamste bron voor zijn) geen informatie over technische velden binnen de bron (bijvoorbeeld gegenereerde primaire sleutels), terwijl deze velden wél in het DIM verwerkt moeten worden, en dus opgenomen moeten zijn in de stuur-metadata. Anderzijds bevatten de FUGEM's (en daardoor IGC) ook informatie over bron-tabellen en -attributen en tabellen die niet (door die bron) aan het DIM geleverd worden, en dus ook niet relevant zijn voor de stuur-metadata.

Rationale maatwerk-beheer voor stuur-metadata

De Record LayOut (RLO) speelt een belangrijke rol bij het vergaren van zowel technische als functionele metadata:

- **Technisch**
  De RLO beschrijft welke gegevens door de bron geleverd worden, en hoe deze gegevens zich historisch gedragen. Deze metadata is onmisbaar voor correcte verwerking en opslag van de gegevens in het Data Integratie Magazijn (DIM).
- **Functioneel:**
  De RLO beschrijft hoe de geleverde gegevens functioneel geïnterpreteerd moeten worden. Deze metadata is randvoorwaardelijk voor het correct verwerken in informatieproducten. Deze categorie is vooral van belang voor de afnemers.

De technische metadata in de RLO zal veelal gebruikt worden om de stuur-metadata te vullen. De functionele metadata zal als basis dienen voor het datamodel in IDA (inclusief verticale lineage) en zal ook als zodanig gekoppeld worden binnen IGC.

Het is lastig om vanuit de RLO geautomatiseerd de stuur-metadata te vullen. Er is bij het vullen van de stuur-metadata namelijk interpretatie van de RLO nodig om het als geheel bruikbaar te laten worden. Denk hierbij bijvoorbeeld aan het beschikbaar hebben van afgeleide velden; deze komen niet direct voor in de RLO, maar dienen wel in de stuur-metadata terecht te komen.

Om dergelijke missende zaken in de RLO geautomatiseerd over te nemen in de stuur-metadata, zouden we daarom eerst de RLO handmatig moeten aanpassen, om deze vervolgens geautomatiseerd over te kunnen nemen.

De RLO is echter eigendom van (en wordt beheerd door) de bron. Daarom zouden deze aanpassingen alleen in een, bij het DIM beheerde, kopie mogen worden uitgevoerd, wat weer zou betekenen dat de aanpassingen bij elke nieuwe versie van de RLO opnieuw (op de kopie) moeten worden doorgevoerd.

Delen van de RLO komen ook in IGC terecht. Inzet van IGC als "tussenstation" tussen RLO en stuur-metadata was dus de moeite van een onderzoek waard. Helaas bleek uit dat onderzoek dat IGC-metadata en stuur-metadata dusdanig verschillen qua structuur en opslagmethode dat inzet van IGC als tussenstation geen haalbare kaart is.

Voor projectfase 4/5 blijft het beheer van (dit deel van) de stuur-metadata een handmatig proces. Op een later tijdstip, als er meer duidelijk is over de, buiten het project, ingerichte beheer-processen voor het DIM, zal opnieuw gekeken worden naar optimalisatie/automatisering van het beheer van stuur-metadata.

Bijlage D: VOLGEN ADMINISTRATIEVE HISTORIE BRON

Deze bijlage wordt in de volgende [GL(31]versie van dit ontwerp ingevuld, en zal een samenvatting bevatten van een aparte, meer gedetailleerde, uitwerking van de behandeling van administratieve tijdlijnen en geldigheidstijdlijnen.

Bijlage E: OVERZICHT STUUR-METADATA

Stuur-metadata m.b.t. de verwerking van bronleveringen

Deze stuur-metadata is gebaseerd op de ~~GLO~~ GIA (o.a. contract-metadata, bewaartermijnen), de RLO (o.a. tabellen, velden, indices) en op algemene maskeringsregels.

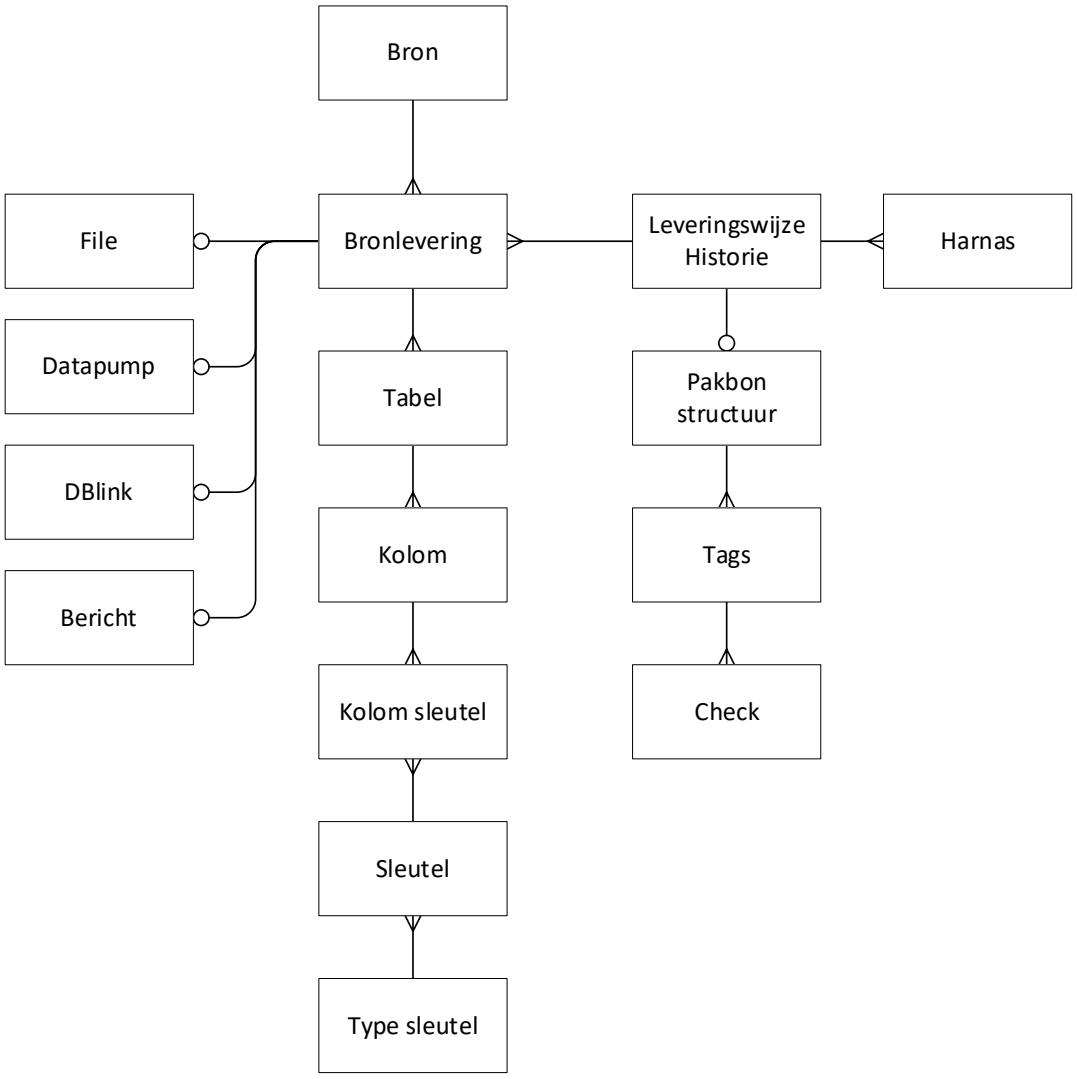
Het doel van deze stuur-metadata is het sturen van de werking van de verschillende harnassen bij het verwerken van de bronlevering vanaf de ontsluitingszone tot en met de bronzone. Ook kunnen de ontkoppelviews op basis hiervan gegenereerd en onderhouden worden.

Voor de verwerking van bronleveringen naar de bronzone worden de volgende blokken stuur-metadata gebruikt:

- Bronlevering (algemeen)
  - Leveringswijze historie
  - Locatie
  - Filemask naam pakbon
  - Filemask naam parfile
  - Filemask naam logfile
  - Filemask data bestanden
  - Leveringstijdvenster

- Bronlevering (inhoud)
  - Tabellen in de bronlevering, met per tabel
    - Leveringswijze historie (incrementeel/stapelbaar/volledig)
  - Kolommen (per tabel), met per kolom:
    - Technisch formaat
    - Identificerend Y/N
    - Maskeringswijze
    - Gebruikt binnen administratieve tijdslijn Y/N
  - Afgeleide velden (per tabel), met, naast bovenstaande, extra per veld:
    - Afleiding (bv. te gebruiken uniformeringslogica)
  - Sleutels (per tabel):
    - Primaire sleutel
    - Eventuele andere (bedrijfs)sleutels
  - Te gebruiken harnassen en/of maatwerk-ETL, zowel voor de verwerking naar de staging-laag als voor die van staging-laag naar bronzone (per tabel)
    - Te gebruiken harnas-versie
    - Binnen het harnas uit te voeren controles
    - Eventueel door het harnas op te starten “nabranders”

Het volgende diagram is een weergave van de voornaamste blokken stuur-metadata en hun onderlinge samenhang.



Voor meer details wordt verwezen naar de documentatie welke opgesteld is/wordt tijdens de realisatie.

Stuur-metadata m.b.t. informatiegebieden en -producten

Metadata voor het aanmaken/verversen van informatiegebieden en informatieproducten. Minder "rijk" dan de vorige categorie, omdat er minder generieke oplossingen worden toegepast.

Deze stuur-metadata wordt eind fase 4 ontworpen. Deze paragraaf [GL(32]zal dan ook in een latere versie van dit ontwerp worden ingevuld.

Stuur-metadata m.b.t. Data Lifecycle Management

Deze stuur-metadata wordt in fase 5 ontworpen. Deze paragraaf [GL(33]zal dan ook in een latere versie van dit ontwerp worden ingevuld.

Bijlage F: GEBRUIK HASH FUNCTIES

Bij het gebruik van hash functies (zoals in DataVault2) gelden een aantal technische randvoorwaarden om te zorgen dat er geen beveiligingsproblemen optreden. Een aantal zijn algemeen en een aantal specifiek.

Algemene randvoorwaarden

Voldoende zware hash functie

Een hash functie dient voldoende veilig te zijn om te zorgen dat de inspanning die benodigd is om een hash te ontsleutelen dusdanig hoog is dat dit "onmogelijk" is. Daarom wordt aangeraden minimaal SHA-256 of EAS-256 als hash-methodiek te gebruiken.

Scheidingsteken gebruiken bij hashes over meerdere velden

Indien een hash over meerdere velden wordt gedaan moet ervoor gezorgd worden dat de velden aan elkaar gekoppeld worden met een scheidingsteken. Dit om te voorkomen dat twee velden met verschillende inhoud toch hetzelfde hash resultaat opleveren.

Verklarend voorbeeld:
Het veld X1 heeft waarde "ABC" en het veld Y1 heeft waarde "DEF".
Zonder scheidingsteken zou het resultaat van het gecombineerde veld een hash van "ABCDEF" zijn. Maar dat zou hetzelfde resultaat zijn voor het geval X2 de waarde "ABCD" en Y2 de waarde "EF" heeft. Om dit te voorkomen zou je de hash van "ABC|DEF" moeten nemen zodat het tweede geval de hash van "ABCD|EF" worden. In het DIM is de pipe ("|") het gekozen scheidingsteken.

Schoning van data voor het hashen

Data die gehasht moet worden zal eerst door een proces moeten waar de informatie geschoond wordt. Voorbeelden hiervan zijn het verwijderen van overbodige spaties, typecasting, standaardisering van gebruik van hoofdletters en corrigeren van titel afkortingen. De geschoonde data zal ook worden opgeslagen in de ongemaskerde satelliet. Indien dit niet gedaan wordt bestaat het risico dat gegevens over verschillende bronnen niet met elkaar gejoined of vergeleken kunnen worden.

Altijd een salt gebruiken als deel van de te hashen sleutel

Om te voorkomen dat er een vertaaltabel gemaakt kan worden zal er aan de hash altijd een salt worden toegevoegd. Een salt is een random string die na generatie ergens beveiligd wordt opgeslagen (in een systeemtabel), en wordt toegevoegd aan de te hashen waarde om extra randomisatie toe te voegen aan een te versleutelen veld. Deze randomisatie is benodigd om te voorkomen dat mensen vertaaltabellen proberen te genereren.

N.B. Elke omgeving (OTAP) heeft zijn eigen set salt's.

Specifieke randvoorwaarden

In de bronzone wordt op twee specifieke plaatsen een hash gebruikt:

- Bij het genereren van de hub-sleutel.
- Bij het maskeren van identificerende gegevens, zoals opgeslagen in de ID-M satelliet, en dan met name bij de maskeringsmethode "koppelbaar vervangen".

Soms is de primaire sleutel van een brontabel een bedrijfssleutel die ook als (gemaskeerd)attribuut in de bronzone wordt opgeslagen, bijvoorbeeld een BSN.

De hash die gebruikt wordt om de DIM-sleutel te genereren dient dan een andere salt te gebruiken dan de hash die gebruikt wordt om, in dit voorbeeld, de BSN te maskeren, zodat DIM-sleutel en gemaskeerd onmogelijk tot dezelfde hash kunnen leiden. Dit is om te zorgen dat er geen koppelingen mogelijk zijn tussen de technische sleutel en de gemaskeerde gegevens.

In zijn algemeenheid geldt:
Om te voorkomen dat gehashte technische velden (sleutels hub en satellieten) en gehashte functionele velden vergeleken kunnen worden dient er voor de twee soorten velden altijd een verschillende salt te worden gebruikt. Deze salts zullen in een technische systeemtabel die niet benaderbaar is door gebruikers worden opgeslagen. Deze technische tabel zal dus alleen toegankelijk voor (de technische user accounts gebruikt door) ETL-programmatuur.

De hash voor het technische veld ziet er dan in pseudocode zo uit:
Hashfunction("<SaltTech>|<bedrijfssleutel>")

En de hash voor het functionele veld zo:
Hashfunction("<SaltFunc>|< bedrijfssleutel >")

Bijlage G: DATAKWALITEITSBEHEER

De verworven ETL+-tooling bevat ook oplossingen voor datakwaliteitsbeheer: QualityStage en IBM Information Analyzer.

Het inrichten van deze tooling is echter geen onderdeel van de fasen 4 of 5, en valt potentieel zelfs volledig buiten scope van het project.

Wel worden er in (de processen rondom) het DIM al uitgangspunten gehanteerd waarop ook de latere uitrol van datakwaliteitsbeheer (incl tooling) zal zijn gebaseerd.

Alles laden, tenzij

Het DIM "weigert niet aan de poort", ook gegevens die qua inhoud afwijken van hun definitie (of van de werkelijkheid) worden geladen.
Het bestaan van foutieve informatie (in de bron) is immers een "administratieve waarheid", en zal dus gewoon in het DIM geladen moeten worden (en daar beschikbaar blijven) totdat de bewaartermijn ervoor verstreken is.

Het DIM bevat een (historische) replica van de bron. Eventuele gegevensreparaties in die bron zullen dus ook in het DIM worden doorgevoerd, zij met enige vertraging.[120]
Deze reparatie wordt in het DIM overigens verwerkt als een nieuwe versie van het gegeven, en niet door de oorspronkelijke (foutieve) gegevens te overschrijven.

Alleen als de data in een bronlevering dusdanig onbetrouwbaar is dat deze de integriteit van het DIM zelf bedreigt, bijvoorbeeld omdat de primaire sleutel niet uniek blijkt te zijn, wordt deze niet ingelezen. Deze "weigering" gebeurt overigens op leveringsniveau: bij één fout record wordt de hele levering geweigerd.
Een fout in één record (of een beperkt aantal records) is namelijk vaak een indicatie van een algemener probleem, meestal niet in de brondata zelf, maar in het aanmaken van de bronlevering (door de bron).

Meten o.b.v. behoefte

Het DIM biedt mogelijkheden om datakwaliteit te meten/toetsen, zowel incidenteel als binnen het reguliere proces. De verantwoordelijkheid voor datakwaliteit blijft echter bij de bron.

De uit te voeren metingen (en de vereiste rapportage daarvan) worden ingericht o.b.v. behoefte (zoals gedefinieerd in requirements van een afnemer), en niet o.b.v. aanbod (zoals gedefinieerd in de RLO).
Dit enerzijds omdat overbodige datakwaliteitsmetingen onnodige rekenkracht én onnodige beheerinspanning vereisen, en anderzijds omdat meldingen van échte datakwaliteitsproblemen (op "critical data elements") anders het risico lopen onder te sneeuwen in minder belangrijke meldingen.

Vaak zal die afnemer een "gewone" afnemer van DIM-informatieproducten zijn, maar ook de bron zelf kan afnemer zijn, bijvoorbeeld als datakwaliteitsmetingen op de bron-replica in het DIM eenvoudiger zijn in te richten dan datakwaliteitsmetingen in de bron zelf.

[120] Deze vertraging is maximaal de periode tussen twee bronleveringen

Overschrijven kan niet, verbeteren wel

Soms blijken gegevens een te lage kwaliteit te hebben om bruikbaar te zijn in informatieproducten, en blijkt verhoging van die kwaliteit in de bron niet (tijdig) mogelijk te zijn.

Mocht dat het geval zijn, dan zijn twee opties om, binnen het informatieproduct (in de bedrijfszone) of in een achterliggend informatiegebied (in de integratiezone), de kwaliteit van de geleverde informatie te verhogen:

- **Door een algoritme**
  - Er wordt een afgeleid veld gedefinieerd waarin, bijvoorbeeld via defaulting of het afvlakken van outliers, een bruikbaarder alternatief voor het oorspronkelijke gegeven beschikbaar komt.
  - Dit afgeleide veld wordt gebruikt in het informatieproduct, maar het oorspronkelijke veld blijft ook beschikbaar. De traceerbaarheid blijft dus in stand.
  - De afnemer is eigenaar van het algoritme; hij/zij specificeert, en accepteert (middels een GAT) de implementatie ervan in het DIM.

- **Door een bron van verbeteringen**
  - Er wordt, buiten het DIM, en door (of in opdracht van) de afnemer, een extra applicatie (de "reparatiebron") gebouwd, waarin verbeteringen van foutieve data staan. Deze applicatie is meestal zeer eenvoudig; een (goed beveiligd) spreadsheet voldoet vaak al.
  - De afnemer controleert de gegevens in het DIM, en waar reparaties noodzakelijke zijn voert hij/zij de verbeteringen, middels een meestal handmatig proces, (door of bij de afnemer) in de reparatiebron.
    De afnemer is er zelf voor verantwoordelijk dat dit invoerproces robuust en betrouwbaar is.
  - Voor het DIM is de reparatiebron "een bron als alle andere", en de gegevenslevering uit die bron (met daarin de reparaties) wordt dus volgens het reguliere proces ingelezen in de bronzone.
  - Er wordt (in integratiezone of bedrijfszone) een afgeleid veld (of een als zodanig herkenbaar afgeleid record) gedefinieerd waarin de waarde uit de "echte" bron wordt gekopieerd, tenzij er, voor dezelfde sleutel, ook een waarde is geleverd door de reparatiebron: dan krijgt het veld/record de waarde uit de reparatiebron.
  - Dit afgeleide veld/record wordt gebruikt in het informatieproduct, maar het oorspronkelijke veld/record blijft ook beschikbaar. De traceerbaarheid blijft dus in stand.
  - De afnemer is eigenaar van de keuzeregels tussen echte bron en reparatiebron; hij/zij specificeert, en accepteert (middels een GAT) de implementatie ervan in het DIM.

Een "bron van verbeteringen" (i.p.v., directe reparatie in het DIM) wordt, bijvoorbeeld, gebruikt voor de SUAG-levering.

Bijlage H: VERSIEBEHEER & RELEASEMANAGEMENT

De functionaliteit van het DIM is onder te verdelen in:

- Database-objecten (tabellen, views, etc)
- DataStage ETL-jobs
- SQL-scripts (m.n. voor reparaties na calamiteiten, zo min mogelijk)[121]
- Linux shell-scripts (voor operating system handelingen welke niet/erg moeilijk door DataStage uitgevoerd kunnen worden)
- Stuur-metadata (bijvoorbeeld voor de harnassen)
- IWS-jobs en -streams

Voor elk van deze categorieën wordt het versiebeheer uitgevoerd door GIT:

- Voor database-objecten o.b.v. de DDL-scripts
- Voor ETL-jobs o.b.v. een code export uit de DataStage-repository, die m.b.v. een Python-script in individuele componenten omgezet is, daarnaast de daarvoor gebruikte datamodellen in IDA
- Voor SQL- en Linux-scripts de scripts zelf
- Voor stuur-metadata moet de basis voor versiebeheer nog bepaald worden

Releases bevatten i.h.a. componenten uit meerdere van bovenstaande categorieën.

Uitrol van een release gebeurt middels door de DWH-DBA's uitgevoerde scripts (en dus niet via XLdeploy). Het DIM sluit hierbij aan op de bestaande praktijk binnen de afdeling DWH.

Deze release-scripts zijn integraal onderdeel van de release, en worden voor elke stap binnen een OTAP-promotie (of tenminste voor de promotie van T naar A, en die van A naar P) ongewijzigd herbruikt.

N.B. Het is mogelijk dat de release-processen en -technieken bij de overgang naar DXC moeten worden aangepast.

Een aandachtspunt is het deployment van informatiegebieden/informatieproducten met twee varianten (gemaskeerd en ongemaskeerd), en van informatieproducten met en zonder zeer gevoelige personen/zaken (bv. VIP's).
De aanpak [GL(34]hiervoor zal in een later stadium (eind fase 4 of fase 5) worden bepaald, als meer bekend is over de precieze opzet van dergelijke deliverables.

[121] Het heeft de voorkeur om zelfs reparaties na calamiteiten via de "formele route" (via harnassen en ontkoppelviews) te laten verlopen. Dit garandeert namelijk dat alle logging wordt doorlopen, en alle technische velden worden aangemaakt.

Bijlage I: CONCEPTEN IDENTIFICEREN/MASKEREN

Het minimaliseren van privacy-risico's is een belangrijk aandachtspunt bij de opslag en verwerking van gegevens in het DIM. Deze bijlage beschrijft de belangrijkste concepten die hierbij gebruikt worden.

N.B. De gebruikte voorbeelden mogen niet als "specificatie" worden gezien; die staat in **Datafabriek - Bescherming van persoonsgegevens in het DIM**, met een verdere uitwerking in **Bescherming van persoonsgegevens in het DIM - Matrix Maskeringsklassen**.

Persoonsgegevens

Gegevens die, direct of indirect, herleidbaar zijn naar een persoon noemen we persoonsgegevens. Op internet kom je hiervoor ook wel de term PII (Personally Identifiable Information) voor tegen.

De Autoriteit Persoonsgegevens geeft als definitie voor een persoonsgegeven:

> *Alle informatie over een geïdentificeerde of identificeerbare natuurlijke persoon. Dit betekent dat informatie ofwel direct over iemand gaat, ofwel naar deze persoon te herleiden is. Gegevens van overleden personen of van organisaties zijn geen persoonsgegevens volgens de AVG.*

De overgrote meerderheid van de gegevens binnen UWV (en daarmee ook binnen het DIM) is dus, gezien vanuit de AVG, een persoonsgegeven.

Binnen het DIM gaan we verschillend om met persoonsgegevens van burgers en persoonsgegevens van UWV-medewerkers:

- Beveiliging van persoonsgegevens van **burgers** zit in het hart van het DIM; die gegevens maskeren we, bijvoorbeeld, al bij opslag. Zie daarvoor de drie stappen in paragraaf 5.1 (Minimalisatie privacy-risico's in 3 stappen).
  Ook bij anonimisering van testdata (buiten het DIM, door TSC) ligt de focus op het onherkenbaar maken van de burger.
- Beveiliging van persoonsgegevens van **medewerkers** gebeurt op een aantal verschillende manieren:
  - **Niet opslaan**
    Gevoelige gegevens (salarissen e.d.) slaan we gewoon niet op. De afdeling HRM is hier de poortwachter. Zij geven (terecht) slechts zeer beperkt toestemming voor het opslaan/gebruiken van medewerkersgegevens in analyses/rapportages, en dus voor levering van die gegevens aan het DIM.
    Als de gegevens wél gebruikt/geleverd mogen worden, dan stelt HRM hier vaak strikte gebruiksbeperkingen aan. Een PUIK[122], bijvoorbeeld, mag wel gebruikt worden om gegevens aan afdelingen te koppelen, maar niet om de activiteiten van individuele medewerkers te analyseren.
  - **Beschouwen als als (bijzondere) burger**
    Als een medewerker ook als burger een relatie heeft met het UWV (hij/zij heeft een uitkering, bijvoorbeeld), dan worden die gegevens binnen het DIM net zo verwerkt als die van "gewone" burgers.
    Als een zaak, omdat die betrekking heeft op (familie van) een medewerker, binnen UWV op een aparte manier of op een apart kantoor wordt afgehandeld, dan zien we dat ook terug in de aan het DIM geleverde gegevens. We kunnen die gegevens

[122] Uniek ID voor elke UWV-medewerker (intern of extern)

dan, op basis van die gegevens, uit de informatieproducten filteren, zodat ze niet zichtbaar zijn voor (niet-geautoriseerde) afnemers.

- **Beschouwen als behandelend medewerker**
  De bulk van de medewerkers-informatie in het DIM is niet veel meer dan het medewerkers-ID (de PUIK) van de behandelend beambte en/of invoerder van gegevens.
  Die PUIK wordt in het DIM ongewijzigd opgeslagen en verwerkt, en dus niet gemaskeerd, zoals we dat wel doen voor, bijvoorbeeld, BSN's van burgers.
  Wel wordt voor elk informatieproduct gekeken of het opnemen van PUIK's wel een doel dient, en of dat een doel is waarvoor HRM toestemming heeft gegeven.

Anoniem en pseudoniem

Anonieme gegevens zijn niet meer te herleiden naar een persoon (en dus geen persoonsgegevens meer), pseudonieme gegevens zijn wel te herleiden naar een persoon, maar je moet er óf de nodig moeite voor doen, óf je kunt het alleen als je speciaal bent geautoriseerd om zulks te doen, bijvoorbeeld omdat je toegang hebt gekregen tot een decodeer-functie.

Gezien de gegevensrijkdom van UWV worden persoonsgegevens pas echt anoniem als je ze verhaspeld (zoals bij testdata), of als je aggregeert.
Helaas maakt verhaspelen de data onbruikbaar voor analyse/rapportage, en kun je aggregatie eigenlijk alleen maar toepassen als je al weet voor welke analyse/rapportage de gegevens precies bedoeld zijn. En dat weet je in het DIM meestal niet vooraf.

Een voorbeeld:

Originele gegevens

Naam	Woonplaats	Beroep	Gezondheid
Jan	Amsterdam	Koning	Goed
Piet	Amsterdam	Keizer	Slecht
Klaas	Amsterdam	Admiraal	Goed
Erik	Amsterdam	Trainer	Goed
Roger	Eindhoven	Trainer	Goed
Rogério	Eindhoven	Prins	Goed

Alle gegevens herkenbaar, je ziet meteen dat de gezondheid van Piet slecht is.

1 - Alleen de sleutel (in dit geval de naam) onherkenbaar gemaakt

Naam	Woonplaats	Beroep	Gezondheid
XYZ	Amsterdam	Koning	Goed
ABC	Amsterdam	Keizer	Slecht
YZA	Amsterdam	Admiraal	Goed
BCD	Amsterdam	Trainer	Goed
ZAB	Eindhoven	Trainer	Goed
CDE	Eindhoven	Prins	Goed

Gegevens niet meer direct herkenbaar, maar als je weet dat Piet Keizer is te Amsterdam kun je achterhalen dat de gezondheid van Piet slecht is.

De gegevens zijn pseudoniem.

2 – Daarnaast ook de woonplaats verhaspeld

Naam	Woonplaats	Beroep	Gezondheid
XYZ	Eindhoven	Koning	Goed
ABC	Eindhoven	Keizer	Slecht
YZA	Eindhoven	Admiraal	Goed
BCD	Eindhoven	Trainer	Goed
ZAB	Amsterdam	Trainer	Goed
CDE	Amsterdam	Prins	Goed

Piet is nu niet meer te vinden (gegevens zijn anoniem), maar de gegevens zijn ook niet meer bruikbaar in analyses. Je kunt, bijvoorbeeld, geen correct percentage ongezonde mensen per woonplaats meer bepalen.

3 – Aggregaat op deel van de attributen

Woonplaats	Gezondheid	Aantal
Amsterdam	Goed	3
Amsterdam	Slecht	1
Eindhoven	Goed	2
Eindhoven	Slecht	0

De gegevens zijn anoniem, maar niet meer bruikbaar in analyses. Je kunt, bijvoorbeeld, niet meer zien dat het aantal trainers in Amsterdam en Eindhoven gelijk is.

4 – "Draaitabel" – Alle mogelijke aggregaten gecombineerd

Woonplaats	Beroep	Gezondheid	Aantal
Amsterdam	Koning	Goed	1
Amsterdam	Koning	Slecht	0
Amsterdam	Keizer	Goed	0
Amsterdam	Keizer	Slecht	1
Amsterdam	Admiraal	Goed	1
Amsterdam	Admiraal	Slecht	0
Amsterdam	Trainer	Goed	1
Amsterdam	Trainer	Slecht	0
Amsterdam	Prins	Goed	0
Amsterdam	Prins	Slecht	0
Eindhoven	Koning	Goed	0
Eindhoven	Koning	Slecht	0
Eindhoven	Keizer	Goed	0
Eindhoven	Keizer	Slecht	0
Eindhoven	Admiraal	Goed	0
Eindhoven	Admiraal	Slecht	0
Eindhoven	Trainer	Goed	1
Eindhoven	Trainer	Slecht	0
Eindhoven	Prins	Goed	1
Eindhoven	Prins	Slecht	0

Je kunt nu weer (bijna) alle analyses aan , maar de tabel explodeert.
Daarnaast zijn de gegevens toch weer, voor combinaties met lage aantallen, herleidbaar geworden naar personen.

In het DIM kiezen we voor optie 2 als we gegevens maskeren in de bronzone, en, waar mogelijk, voor optie 3 als we gegevens op maat maken in de bedrijfszone (zie "**Fout! Verwijzingsbron niet gevonden.**~~Gegevensminimalisatie bij levering~~" later in deze bijlage).

De **UWV ICT Richtlijn Data Anonimisering en Pseudonimisering Versie 1.1** bevat goede achtergrondinformatie over anonimiseren en pseudonimiseren.
Let op: Via de Digitale Werkplek vind je alleen een 1.0-versie. Die is achterhaald, en gaat vooral over testdata.

Identificerende en niet-identificerende attributen

Conform de AVG-definitie is een persoonsgegeven informatie die ofwel direct over iemand gaat, ofwel naar deze persoon te herleiden is.

Binnen het DIM classificeren we attributen waarlangs je een gegeven (vrijwel) direct naar een persoon kunt herleiden als "identificerend".

We gebruiken die classificatie vooral om handig om te kunnen gaan met gegevensmaskering, en het splitsen van gegevens in "gemaskeerd" en "ongemaskeerd". Het is dus géén officiële term uit de AVG (of uit andere wet- of regelgeving). Wel hebben we met de beleidsafdelingen van UWV (BSO, CISO, bureau FG) afgestemd welke gegevens we als identificerend gaan behandelen, en welke niet.

Twee voorbeelden:

1. **We krijgen uit een bron een tabel "uitkeringsrecht" met daarin BSN, ingangsdatum, einddatum, type uitkeringsrecht.**
   Alleen de BSN verwijst daarbij naar een persoon, en is dus een identificerend attribuut.
   De overige attributen beschrijven het uitkeringsrecht, en niet de persoon, en zijn dus niet-identificerend.
2. **We krijgen uit een bron een tabel "baan" met daarin BSN, KvK-nummer werkgever, ingangsdatum, einddatum, functie, aantal uren.**
   Naast de BSN is nu ook het KvK-nummer identificerend, omdat, voor eenmanszaken en andere kleine ondernemingen, de persoon eenvoudig uit het bedrijf kunt afleiden.
   De overige attributen zijn weer niet-identificerend.

In beide gevallen zou je overigens, via de niet-identificerende attributen, de gegevens nog naar een persoon kunnen herleiden, zeker als je nog wat context hebt, en voor bijzonder personen. Het aantal circusdirecteuren dat 2 jaar in functie is zal, bijvoorbeeld, niet heel groot zijn.

Maar zoals gezegd, de classificatie "identificerend" gebruiken we vooral om vast te leggen welke attributen we onherkenbaar(der) moeten maken.

Belangrijk:
Alleen de eenvoud waarmee je gegevens via een attribuut zou kunnen herleiden naar een persoon bepaalt of dat attribuut als "identificerend" moet worden geclassificeerd; de vertrouwelijkheid van de gegevens zelf is daarop niet van invloed. We classificeren dus, bijvoorbeeld, een diagnosecode niet als identificerend; de code zélf is immers niet identificerend.

Maskeren van identificerende gegevens (bij opslag)

In de bronzone splitsen we de binnengekomen data, op basis van de classificatie in de stuur-metadata (die zelf weer uit de RLO's komt, en dus in overleg met de bronnen tot stand is gekomen) in "identificerende" en "niet-identificerende gegevens", en voorzien we alle identificerende attributen van een gemaskeerd alternatief, conform de eisen die daar in de stuur-metadata voor zijn vastgelegd.

De maskeringsregels zijn (voorlopig) DIM-specifiek. We hebben wel weer met de beleidsafdelingen van UWV afgestemd hoe we welk gegeven gaan maskeren.

Belangrijk:
Als hetzelfde identificerende attribuut in meerdere tabellen voorkomt maskeren we dat altijd volgens dezelfde regels, en ook hier is de vertrouwelijkheid van de gegevens daarop niet van invloed. We maskeren dus, bijvoorbeeld, het KVK-nummer bij Baan (in het voorbeeld hierboven) op dezelfde wijze als het KVK-nummer in een Werkgevers-tabel (strikt genomen géén persoonsgegeven). Dit doen we zowel om de gegevens referentieel integer te houden, als om te voorkomen dat gevoelige gegevens, door ze te koppelen met minder goed gemaskeerde ongevoelige gegevens, alsnog eenvoudig te herleiden zijn naar een persoon.
Voor de voorbeelden uit de vorige paragraaf:

Uitkeringsrecht		
Identificerend (ongemaskeerd)	**Identificerend (gemaskeerd)**	**Niet-identificerend**
BSN (ongemaskeerd)	BSN (gemaskeerd)	ingangsdatum einddatum type uitkeringsrecht

Baan		
Identificerend (ongemaskeerd)	**Identificerend (gemaskeerd)**	**Niet-identificerend**
BSN (ongemaskeerd) KvK-nummer (ongemaskeerd)	BSN (gemaskeerd) KvK-nummer (gemaskeerd)	ingangsdatum einddatum functie aantal uren

Extra voorbeeldje: een tabel "UWV-afdeling", met daarin afdelingscode, afdelingsnaam, peildatum, aantal medewerkers:

UWV-afdeling		
Identificerend (ongemaskeerd)	**Identificerend (gemaskeerd)**	**Niet-identificerend**
niets…	*nog steeds niets…*	afdelingscode afdelingsnaam~~code~~ peildatum aantal medewerkers

Er staan nu dus in de bronzone drie groepen attributen per brontabel (in de "satellieten", waarvan er als de brontabel helemaal geen persoonsgegevens bevatte, twee leeg zijn).

N.B. In bijzondere gevallen is een attribuut geclassificeerd als "identificerend", maar is de uitgevoerde maskering "kopieer ongewijzigd". De gemaskeerde en de ongemaskeerde waarde zijn dan dus aan elkaar gelijk. Een voorbeeld hiervan is de geboortedatum. Dit veld is identificerend, maar ook een belangrijke input voor de bedrijfsregels van UWV. Maskeren van ~~de~~ die datum (tot geboortejaar of geboortemaand) zou de waarde voor analyse/rapportage dus sterk verminderen.
Om governance-redenen (expliciet maken van bovenstaande afweging) wordt een dergelijk attribuut toch als "identificerend" geclassificeerd en verwerkt.

Van drie groepen attributen naar drie typen gegevens

De driedeling in de bronzone (2x identificerend en 1x niet-identificerend) is handig voor de verwerking in de bronzone, maar niet handig in het gebruik. Ook de governance en de toegangsregels zijn afhankelijk van de gevoeligheid van de totale gegevensset, dus inclusief niet-identificerende attributen.

Vandaar dat de rest van het DIM gebruik maakt van een andere driedeling[GL(35)]:

- Ongemaskeerde persoonsgegevens
- Gemaskeerde persoonsgegevens
- Niet-persoonsgegevens (ook wel "overige gegevens)

De bronzone-ontkoppelviews maken dit mogelijk door de drie "satellieten" uit de bronzone steeds handig te combineren.

De voorbeelden in de vorige paragraaf resulteren dan in de volgende ontkoppelviews:

Uitkeringsrecht

Ongemaskeerde persoonsgegevens	Gemaskeerde persoonsgegevens	Niet-persoonsgegevens
BSN (ongemaskeerd)	BSN (gemaskeerd)	*(niet van toepassing)*
ingangsdatum einddatum type uitkeringsrecht	ingangsdatum einddatum type uitkeringsrecht	

Baan

Ongemaskeerde persoonsgegevens	Gemaskeerde persoonsgegevens	Niet-persoonsgegevens
BSN (ongemaskeerd) KvK-nummer (ongemaskeerd	BSN (gemaskeerd) KvK-nummer (gemaskeerd)	*(niet van toepassing)*
ingangsdatum einddatum functie aantal uren	ingangsdatum einddatum functie aantal uren	

UWV-afdeling

Ongemaskeerde persoonsgegevens	Gemaskeerde persoonsgegevens	Niet-persoonsgegevens
(niet van toepassing)	*(niet van toepassing)*	afdelingscode afdelingsnaam~~code~~ peildatum aantal medewerkers

Drie typen gegevens voor twee groepen afnemers

Het DIM heeft twee groepen afnemers (eigenlijk: afnemende processen):

- **Sturen & Verantwoorden**
  Management informatie kan (vrijwel) altijd worden samengesteld op basis van gemaskeerde persoonsgegevens. Het gebruik van ongemaskeerde gegevens voor S&V zou dus onnodige privacy-risico's opleveren, en dat is niet AVG-compliant.
- **Operationeel gebruik**
  Informatieproducten voor operationeel gebruik (bijvoorbeeld de SUAG-levering) kunnen juist (vrijwel) nooit worden samengesteld op basis van gemaskeerde persoonsgegevens. Het gebruik van ongemaskeerde gegevens is hier dus onvermijdbaar, en dus AVG-compliant (mits het operationele proces zelf AVG-compliant is, uiteraard).

In beide gevallen mogen, wat de AVG betreft, ook niet persoonsgegevens worden gebruikt.

Combineren van gemaskeerde en ongemaskeerde persoonsgegevens staan we in het DIM niet toe. Via die route zou een afnemer namelijk gemaskeerde gegevens weer kunnen "ontmaskeren", en zo de volledige gegevensbescherming via maskering tenietdoen.

In de rest van de keten (via de integratiezone naar de informatieproducten in de bedrijfszone) houden we dus de driedeling gemaskeerd/ongemaskeerd in stand. Het toevoegen van niet-persoonsgegevens aan al dan niet gemaskeerde persoonsgegevens is wel toegestaan.
Dat betekent dus:

- Een informatiegebied of informatieproduct voor S&V zal gemaskeerde persoonsgegevens bevatten (vaak gecombineerd met niet-persoonsgegevens)
- Een informatiegebied of informatieproduct voor operationeel gebruik zal ongemaskeerde persoonsgegevens bevatten (weer vaak gecombineerd met niet-persoonsgegevens)

De AVG-termen op dit gebied zijn proportionaliteit (rechtvaardigen de baten van de privacy-risico's) en subsidiariteit (is er geen manier om met minder privacy-risico's hetzelfde te bereiken).

Gegevensminimalisatie bij levering

Gemaskeerd of ongemaskeerd: de laatste stap is dat het geleverde informatieproduct niet meer persoonsgegevens bevat dan noodzakelijk voor het doelproces.

Hierbij wordt, anders dan bij de eerder in deze bijlage beschreven stappen, de vertrouwelijkheid van de gegevens wél uitdrukkelijk meegenomen.

In tegenstelling tot de "brede bronontsluiting" van de bronnen op het DIM is hier dus juist sprake van een "zo smal mogelijk" product: zo min mogelijk attributen, zo min mogelijk rijen.
Filters op VIP's en Eigen Personeel (EP) zijn daarbij een bijzondere vorm van "zo min mogelijk rijen".[123]

De gegevens die wél geleverd worden kunnen nog extra ongevoelig gemaakt worden, bijvoorbeeld door te aggregeren of te classificeren.[124]

Een voorbeeld:

Gedetailleerde gegevens

BSN (gemaskeerd)	Woonplaats	Salaris
ABC	Amsterdam	32.000
DEF	Amsterdam	64.000
GHI	Eindhoven	35.000
JKL	Eindhoven	45.000

Aggregatie

Woonplaats	Aantal salarissen	Totaal salaris
Amsterdam	2	96.000
Eindhoven	2	80.000

Classificatie

BSN (gemaskeerd)	Woonplaats	Salarisklasse
ABC	Amsterdam	30k-40k
DEF	Amsterdam	50k+
GHI	Eindhoven	30k-40k
JKL	Eindhoven	40k-50k

[123] Meer informatie over VIP-, EP- en BZ-filtering in Bijlage J: Filtering op VIP's en BZ/EP+
[124] De volledige lijst staat in **Datafabriek - Bescherming van persoonsgegevens in het DIM**

Misbruik bij de afnemers (“restrisico’s”)

Hoe netjes we het ook bij DIM, DWH en GLV hebben ingericht: uiteindelijk staat of valt de hele AVG-compliance met het gedrag van de afnemers.

Als de afnemers gegevens voor iets anders gebruiken dan ze ons verteld hebben, ze te lang bewaren, of doorleveren aan collega’s, dan levert dat compliance-issues en misschien zelfs datalekken op waar wij, als gegevensleverancier, maar weinig grip op hebben.

Dit zijn de in de GEB beschreven “restrisico’s”.

Het enige dat we, als DIM, DWH en GLV, kunnen doen is goede afspraken maken:

- **Afnemers** in de **SNO/GLA** laten beloven dat ze gegevens **niet** voor iets anders gaan gebruiken dan ze ons verteld hebben, ze **niet** te lang zullen bewaren, en ze **niet** zullen doorleveren aan collega’s.
- **Broneigenaren** in de **GLO/GIA** laten accepteren dat wij er niets aan kunnen doen als een afnemer zich niet aan de SNO/GLA houdt.
- **Zelf** niet in de valkuil van urgente **oplossingen zonder contractuele afdekking** trappen; als een afnemer urgent data nodig heeft, dan hebben ze ook de urgentie om het papierwerk (of een ontheffing daarvan) snel te regelen.

Bijlage J: FILTERING OP VIP'S EN BZ/EP+

Deze bijlage is slechts een introductie; de precieze opzet van de VIP-filtering zal worden beschreven in een apart detailontwerp[G.J.36].

De VIP-tabel

De basis van alle filtering is een door UPA aangeleverde tabel met de BSN's en hun VIP-status.
Op basis van deze levering wordt binnen het DIM een VIP-tabel opgebouwd.
Zowel levering als tabel bevatten een geldigheidstijdlijn. Deze wordt echter alleen t.b.v. de traceerbaarheid bijgehouden: alleen de laatste versie bepaalt of iemand VIP is, en als iemand VIP is geldt dat met terugwerkende kracht.

De VIP-tabel is, net als UPA, opgehangen aan de BSN.
Het kan echter, bijvoorbeeld om VIP-filtering o.b.v. andere persoons-ID's te versnellen en/of te vereenvoudigen, noodzakelijk zijn om de VIP-tabel (of aanvullende filtertabellen) uit te breiden met "alternate keys".

Filteren VIP-gegevens – drie manieren

VIP-gegevens kunnen op drie manieren uit een informatieproduct verwijderd worden:

1. Alleen identificerende rijen van VIP's wegfilteren
2. Alle aan VIP's gerelateerde rijen wegfilteren
3. Aan VIP's gerelateerde identificerende kenmerken verbergen

Alleen identificerende rijen van VIP's wegfilteren

Soms voldoet het om aan VIP's gerelateerde rijen alleen te verwijderen uit de tabellen met identificerende kenmerken.

Voordeel: Eenvoudig te bouwen
Nadeel: Referentiële integriteit kan niet meer gegarandeerd worden, waardoor onverwachte query-resultaten kunnen voorkomen. VIP-gegevens worden namelijk alleen niet meegenomen in totaaltellingen op "child"-tabellen (waar de VIP-rijen dus niet uitgefilterd zijn) als die tabellen gejoind zijn met een "parent"-tabel waar de VIP-rijen wél uitgefilterd zijn.

Alle aan VIP's gerelateerde rijen wegfilteren

Als onverwachte query-resultaten (zie hierboven) niet acceptabel zijn, dan moeten de aan VIP's gerelateerde rijen uit alle tabellen in het informatieproduct worden gefilterd.
Je moet daarvoor óf een soort "cascading filter" bouwen (een rij uit een "child" tabel wordt uitgefilterd als die rij verwijst naar een rij in een "parent" tabel die ook is uitgefilterd.

Voordeel: Eenduidige, consistente dataset
Nadeel: Complex om te bouwen, niet zonder meer toe te passen op ruwe datasets (zoals gegevensvensters direct op de bronzone-ontkoppelviews).

Aan VIP's gerelateerde identificerende kenmerken verbergen

Soms kan een tabel binnen een informatieproduct voldoende "veilig" gemaakt worden door een identificerend veld te "verbergen" (dus leeg te maken, of te vervangen door X'jes o.i.d). Dit is vooral van toepassing als dat identificerende veld niet gebruikt wordt als (verwijs)sleutel.

N.B. Als het veld sowieso niet erg belangrijk is, dan is het natuurlijk het eenvoudigst om het überhaupt niet op te nemen in het informatiegebied (niet voor VIP's maar ook niet voor anderen).

Voordeel: Verbergen (van niet-sleutelvelden) heeft geen impact op de referentiële integriteit
Nadeel: Als het al dan niet moeten verbergen van een veldwaarde niet kan worden afgeleid van de veldwaarde zelf, dan is complexe afleidingslogica (vergelijkbaar met het in de vorige paragraaf~~g~~ beschreven "cascading filter") potentieel vereist.

Filteren van VIP-gegevens – aanpak per type informatieproduct

Onderstaande heeft alleen betrekking op informatieproducten met persoonsgegevens. VIP-filtering is irrelevant voor informatieproducten waarvan de inhoud, bijvoorbeeld door aggregatie, niet meer op personen te herleiden.
Ook voor gemaskeerde informatieproducten is VIP-filtering vrijwel nooit relevant.

VIP-filtering in datamarts

In principe worden in een datamart alle aan VIP's gerelateerde rijen pas bij bevraging uitgefilterd. Dit gebeurt dan door de datamart altijd via een gegevensvenster (met daarin het VIP-filter) te benaderen[125]

Beste optie is om dat te doen door aan alle, in direct of indirect, aan personen gerelateerde afgeleide tabellen binnen de datamart (voor sterschema's dus tenminste de feittabel en de persoonsdimensie) een verwijssleutel naar de VIP-tabel op te nemen.
De VIP-filters in het bovenliggende gegevensvenster vereisen dan alleen een join met die VIP-tabel.

Als dat toch nog onvoldoende performant blijkt te zijn, dan kan aan, direct of indirect, aan personen gerelateerde afgeleide tabellen binnen de datamart een VIP-indicator toegevoegd worden die aangeeft op de betreffende rij direct of indirect op een VIP betrekking heeft. Het filterende gegevensgevenster hoeft nu slechts op de VIP-indicator te filteren. Dit kan direct in de view-definitie; joins zijn daarbij niet noodzakelijk.

Omdat VIP-wijzigingen met terugwerkende kracht gelden, stelt dit wel eisen aan de verversingslogica voor de datamart:

a) Updates op rijen in de datamart tabellen moeten mogelijk zijn
b) Deze updates moeten met relatief hoge frequentie (dagelijks of wekelijks) worden uitgevoerd.

Gelukkig is het aantal VIP's beperkt, en het aantal wijzigingen in het VIP-schap zo mogelijk nog beperkter.

Soms is hiernaast ook noodzakelijk om additionele aan VIP's gerelateerde identificerende kenmerken verbergen. De opties en afwegingen zijn gelijk aan die voor rij-filtering.

VIP-filtering in gegevensvensters die gebaseerd zijn op afgeleide tabellen

Als een gegevensvenster op afgeleide tabellen is gebaseerd (uit een datamart of een informatiegebied), dan gelden de afwegingen voor een datamart.
Zie daarvoor verder de paragraaf hierboven.

VIP-filtering in gegevensvensters die NIET gebaseerd zijn op afgeleide tabellen

Als een gegevensvenster **niet** op afgeleide tabellen is gebaseerd (maar op de bronzone-ontkoppelviews), dan is "alleen identificerende rijen van VIP's wegfilteren" nog eenvoudig uit te

[125] Zie paragraaf 10.4.1.5 (Met en zonder zeer gevoelige gegevens)

voeren door de in de gegevensvenster-view op de betreffende tabellen een join of "where not exists" conditie met de VIP-tabel op te nemen.

Voor "cascading" filters zoals vereist voor "alle aan VIP's gerelateerde rijen wegfilteren" is zijn echter potentieel complexere filter-joins of -condities vereist:

- Voor tijdelijke gegevensvensters (waaronder ook prototypes) en/of gegevensvensters waarvoor de performance-impact van deze complexe filters acceptabel is, kunnen ook deze complexere filter-joins/condities in de gegevensvenster-views worden opgenomen.

- Voor de overige gegevensvensters zijn voorbewerkingen noodzakelijk om de VIP-filters performant te houden. Dit betekent dus dat het gegevensvenster niet puur op de bronzone-ontkoppelviews gebaseerd kan zijn, maar ook deels gebruik moet maken van afgeleide gegevens (uit de integratiezone) waarin een deel van de filter-logica al is uitgevoerd.
  Voorbeelden van benaderingen voor deze afgeleide gegevens zijn:
  - Parallel-tabellen voor de relevante bronzone-ontkoppelviews, enkel bestaande uit een primaire sleutel (gelijk aan die van de ontkoppelview) en een verwijssleutel naar de VIP-tabel.
  - Parallel-tabellen voor de relevante bronzone-ontkoppelviews, enkel bestaande uit een primaire sleutel (gelijk aan die van de ontkoppelview) en de VIP-indicator van de persoon waarop de rij in de ontkoppelview (indirect) betrekking heeft.

  De eerste optie heeft daarbij, net als bij datamarts, de voorkeur.

VIP-filtering in bestandsleveringen

VIP-filtering in bestandsleveringen (indien vereist) moet zo kort mogelijk voor de levering worden uitgevoerd.
Inzet van gegevensvensters voor deze filtering, gezien het "push"-karakter van bestandsleveringen is alleen relevant als de bestandslevering gebaseerd is op andere informatieproducten (bijvoorbeeld datamarts), en er voor die informatieproducten al filterende gegevensvensters beschikbaar zijn.

Filteren van BZ-gegevens (en vergelijkbaar)

Er bestaan plannen om binnen UWV een centrale registratie voor BZ/BGB/EP+ op te zetten. Zodra die er is zal het DIM erop gaan aansluiten, en zal de filterbenadering voor BZ/BGB/EP+ vergelijkbaar worden met die voor VIP's, zij het deels op zaak- i.p.v. persoonsniveau.

Tot die tijd zal het afschermen van BZ/BGB/EP+ grotendeels gebaseerd zijn op gegevens uit de bron zelf.

De toe te passen filterlogica zal dus ook per bron verschillen.

De implementatie van deze BZ-filterlogica kent, in grote lijnen, dezelfde aandachtspunten en mogelijke oplossingen als die van de BZ-filterlogica.
Zie daarvoor dus eerder in deze bijlage.