Whitepaper
GEïntegreerde **R**egionale **D**ata-infrastructuur **A**chterhoek (GERDA)

Een analyse naar de mogelijkheden voor een domeinoverstijgende data-infrastructuur

Datum:
5 juli 2022

Door:
Werkgroep GERDA (Ilona Oude Nijhuis, Lotte van Dieren, Lars Bannink, Maarten den Braber)
I.s.m. Population Health Data NL - als onderdeel van het werkpakket data engineering.

Met hulp van:
Reviews door Freya de Mink (Roseman Labs), Marieke Vermue (ICTU), Martine van de Gaar (Linksight), Peter-Paul Essers (regio Zeeland), Marco Spruit (Leiden University) en André Dekker (Personal Health Train).

Voor:
Domeinoverstijgende samenwerkingsverbanden in (andere) regio's, met name werkgroepen die zich bezig houden met data(analyse).

Over dit document

Contactpersonen: ilonaoudenijhuis@room-to.nl en maarten@populationhealthdata.nl.

Inhoudsopgave

Samenvatting

In Nederland is het vertalen van gezondheidsdata naar inzichten en informatie een uitdaging door de wijze waarop de (preventieve) gezondheidszorg georganiseerd is. Zorgwetten zijn verdeeld over diverse domeinen, met een ander uitvoeringsniveau (landelijk of regionaal) en andere verantwoordelijke en/of uitvoerende partijen. Daarnaast zijn voor een integrale kijk op gezondheid niet alleen medische gegevens relevant, maar ook sociaal-economische gegevens of informatie over de leefomgeving.

Om te komen tot integrale gezondheidsbevordering is het nodig om inzichten op persoonsniveau *over* deze wetten, organisaties en disciplines heen inzichtelijk te maken, door domeinoverstijgend gegevens bij elkaar te brengen. Op dit moment wordt hier in Nederland in verschillende regio's en voor diverse thema's in afzonderlijk onderzoek aan gewerkt. Met dit whitepaper willen 8RHK Gezond en Population Health Data NL de regio's een handreiking doen voor de wijze waarop een geïntegreerde regionale data-infrastructuur voor gezondheid duurzaam kan worden ingericht.

Het inrichten van de data-infrastructuur is maatwerk en kan verschillen per regio, afhankelijk van het beoogde doel en de informatievragen. Tegelijkertijd zijn er een aantal ontwerpkeuzes die kunnen zorgen voor een gemeenschappelijk fundament. Fundamentele ontwerpkeuzes die gemaakt moeten worden gaan over:

1. data-architectuur; centraal versus gefedereerd
2. aggregatieniveau; n=1 versus geaggregeerd niveau
3. benodigde data-integratie; horizontaal onderzoek (bij elkaar brengen van dezelfde data van verschillende personen) versus verticaal onderzoek (bij elkaar brengen van verschillende data van één persoon).

Voor iedere (combinatie van) ontwerpkeuze(s) worden in deze whitepaper bekende en beschikbare oplossingen in kaart gebracht. Al deze oplossingen worden beschreven in termen van en/of getoetst op de aspecten:

- achtergrond
- gebruikte technologie
- data governance
- beveiliging en juridische/privacy aspecten
- ethische en sociale aspecten
- relevantie voor een geïntegreerde regionale data-infrastructuur.

Zes (technische) mogelijkheden die toepasbaar zijn als (onderdeel van) een geïntegreerde regionale data-infrastructuur worden geanalyseerd en beschreven aan de hand van bovenstaande aspecten:

- nationaal datalake
- regionaal datalake
- FAIR data stations
- datavirtualisatie
- Federated Learning (FL)
- Multi Party Computation (MPC)

Op basis van het doel van en de informatievragen voor de GEïntegreerde Regionale Data-infrastructuur Achterhoek (GERDA), vertaald naar een lijst van specificaties, is vervolgens gekeken welke technische oplossing(en) het meest relevant zijn om verder uit te werken in een technisch ontwerp.

Op basis van de uitgangspunten 'data blijft zoveel mogelijk bij de bron' (DIZRA principe) en 'zo min mogelijk handmatige handelingen' (uitgangspunten efficiënt en veilig) wordt voor de basis van de regionale geïntegreerde data-infrastructuur gekozen voor een *gefedereerde data-architectuur*. Omdat de data 'faciliterend moet zijn voor gebruik op persoonsniveau' vanwege 'wederkerig gebruik' voor primair en secundair gebruik (uitgangspunten relevant voor de regio) wordt gekozen voor een gewenst *aggregatievniveau van n=1.* Tot slot wordt, omdat het gaat om analyse van zorgdata (en daarmee bijzondere persoonsgegevens) op n=1 niveau gekozen voor een techniek die uitgaat van 'privacy-by-design' (uitgangspunt efficiënt en veilig; AVG) en die geschikt is voor *verticaal onderzoek*; het samenvoegen van verschillende data van één persoon. Zie voor een overzicht van alle vergeleken oplossingen en aspecten de vergelijkingstabel in hoofdstuk 6.

Multi Party Computation (in combinatie met datavirtualisatie) is daarmee de meest relevante techniek om verder te onderzoeken en mee te nemen in een technisch ontwerp voor GERDA. Daarbij opgemerkt dat binnen de overige oplossingen (o.a. FAIR data stations) elementen zitten die op onderdelen ook kunnen bijdragen aan de gewenste regionale oplossing.

In het volgende werkpakket van de werkgroep GERDA wordt i.s.m. Population Health Data NL uitgewerkt welke gebruikersgroepen, met welke tooling, in de data-werkplaats tot een antwoord kunnen komen. Daarnaast wordt onderzocht of het mogelijk is een *proof of concept* van MPC in te richten rondom acute zorg in de Achterhoek.

Wil je bijdragen aan de (door)ontwikkeling van het technische ontwerp van GERDA, laat het ons dan weten. Neem hiervoor contact op met Ilona Oude Nijhuis (ilonaoudenijhuis@room-to.nl) of Maarten den Braber (maarten@populationhealthdata.nl)

1 Aanleiding en context

1.1 Aanleiding

In Nederland is het omzetten van gezondheidsdata naar inzichten en informatie een uitdaging door de wijze waarop de (preventieve) gezondheidszorg georganiseerd is. Er is sprake van een gefedereerde inrichting met verschillende zorgwetten voor de domeinen. De Wet langdurige zorg (Wlz) heeft een nationale uitvoering. Bij de uitvoering van de Zorgverzekeringswet (Zvw) spelen private zorgverzekeraars een belangrijke rol. De Wet maatschappelijk ondersteuning (Wmo) en de Jeugdwet (JW) zijn er voor de overige vormen van zorg en ondersteuning, terwijl de Wet publieke gezondheid (WPG) de openbare gezondheidszorg regelt; de bestrijding van infectieziekten-crises en de isolatie van personen en/of vervoersmiddelen die internationaal gezondheidsgevaren kunnen opleveren. Ook regelt de WPG de jeugd- en ouderengezondheidszorg. De ongeveer 400 Nederlandse gemeenten zijn primair verantwoordelijk voor de uitvoering van deze laatste drie wetten[1].

Om te komen tot integrale gezondheidsbevordering is het nodig om inzichten van één persoon over deze wetten/organisaties/disciplines inzichtelijk te maken (in een regio-context) door benodigde gegevens bij elkaar te brengen. Dit is mogelijk via een geïntegreerde publieke regionale data-infrastructuur. De regionale data-infrastructuur faciliteert secundair datagebruik (voor o.a. het opstellen van regionaal beleid), waarbij het uiteindelijk mogelijk moet zijn om resultaten terug te koppelen naar het primair proces. Een korte uitwerking van de samenhang en verschillen in het doel, de soort data die gebruikt wordt, de geldende beleidskaders en wat/met wie uitgewisseld kan worden, is uitgewerkt voor primair en secundair datagebruik in Figuur 1.

Figuur 1: Een overzicht van samenhang en verschillen tussen de doelen, kenmerken van de gebruikte data, geldende beleidskaders en mogelijkheden voor data/informatie-uitwisseling voor primair (A: Cliënten en Patiëntenzorg (bron) & Financiële Administratie/verwerking) en secundair (B: Publieke gezondheidszorg, C: Wetenschappelijk onderzoek) gebruik van data.

[1] Zorginstituut Nederland: visiestuk datastrategie en framework gezondheidszorg.

1.2 Context

GERDA is de projectnaam voor het realiseren van een GEïntegreerde Regionale Data-infrastructuur Achterhoek. GERDA heeft als doel invulling te geven aan een data-infrastructuur die data integreert voor primair en secundair gebruik zoals deze infrastructuur in de vorige paragraaf is geïntroduceerd. GERDA is een gecombineerde opdracht van de Vereniging Digitale Zorg Achterhoek en 8RHK Gezond. Dit is in het tekstvak hieronder kort uitgewerkt.

Vereniging Digitale Zorg Achterhoek: Geïntegreerde regionale data-infrastructuur voor beschikbaar stellen en uitwisselen van data vanuit het **primair** proces van zorg (zorgverlener <-> zorgverlener en zorgverlener <-> burger)

+

8RHK Gezond (kavelmodel): Geïntegreerde regionale data-infrastructuur voor de uitwisseling, verwerking, analyse en presentatie voor **secundair** gebruik van data over gezondheid en zorg.

=

GERDA: GEïntegreerde **R**egionale **D**ata-infrastructuur **A**chterhoek voor uitwisseling, verwerking, analyse en presentatie van data over gezondheid en zorg, uit zowel primair proces als secundaire bronnen (bijv. registerdata zoals CBS)

Voor het bepalen van de optimale inrichting van GERDA zijn de doelstellingen van 8RHK Gezond en de Vereniging Digitale Zorg Achterhoek uitgewerkt in onderstaande specificaties:

1. **Aansluitend bij DIZRA principes.** Kaderstellend voor de inrichting van de informatiehuishouding zijn de basisprincipes van DIZRA[2], zoals deze ook onderschreven worden in de Doelarchitectuur Digitale Zorg regio Achterhoek. De belangrijkste zes basisprincipes in relatie tot de regionale data-infrastructuur zijn:
    a. In het informatiestelsel hebben cliënten **regie op hun eigen gezondheidsgegevens** en kunnen deze gegevens meenemen en delen in hun reis door het zorglandschap en in het netwerk van zorgverleners en ondersteuners dat zich rondom hen vormt.
    b. De **data** blijft **bij de bron**, onder de verantwoordelijkheid van de bronhouder, voor een veilig en vertrouwd informatiestelsel waarin het voor cliënten transparant is welke bronhouders welke gezondheidsgegevens registreren en wie het raadpleegt.
    c. Het informatiestelsel hanteert een **gelijk speelveld voor alle leveranciers**. Afspraken worden gemaakt over het gebruik van standaarden, niet over het gebruik van een product of dienst. Iedere organisatie kiest haar eigen leveranciers voor het implementeren van de standaarden.

[2] Manifest - DIZRA

d. Data wordt enkelvoudig geregistreerd bij de bron en vervolgens beschikbaar gesteld voor meervoudig gebruik in verschillende toepassingen. Hiervoor hanteert het informatiestelsel de **FAIR-data** principes.
e. Data is **machine-leesbaar**, machines begrijpen de data, zonder daarbij de leesbaarheid van deze data voor mensen uit het oog te verliezen. Dit opent de mogelijkheden van data-analyse en data-science.
f. In het informatiestelsel wordt **federatief** samengewerkt aan afspraken voor data en voor services. Iedereen implementeert deze afspraken en is aanspreekbaar op het nakomen van de afspraken en de kwaliteit van de implementatie.

2. **Relevant voor de regio.** Analyse domeinoverstijgende data-analyse van data over een persoon in meerdere gekoppelde bronnen ("verticale data-analyse")[3], waarbij de volgende uitgangspunten van belang zijn:
    a. Faciliterend voor gebruik **op persoonsniveau** (import, analyse en output).
    b. Faciliterend voor **wederkerig gebruik** van data voor primaire en secundaire doeleinden.
    c. Focus ligt (vooralsnog) op gestructureerde en ongestructureerde **numerieke en tekstuele data** (en niet op beeld of geluid).
    d. We maken gebruik van **actuele data**, waarbij we verwachten dat data in een hogere frequentie beschikbaar komt in de komende jaren.
    e. We onderkennen de noodzaak voor een **eenduidige betekenis, de herleidbaarheid van gegevens (in de informatiehuishouding)** en **kwaliteit** van beschikbare data.
    f. We ambiëren een **duurzame infrastructuur** - geen inrichting geschikt voor project context.
    g. De datawerkplaats voorziet in **brede mogelijkheden** voor onderzoek en monitoring nu en in de toekomst (brede grond voor data-onderzoek, na afspraken hierover binnen de governance context)

3. **Efficient en veilig.**
    a. Zo **min mogelijk** handmatige handelingen - zowel voor data ingestie & data translatie (Extract, Transform, Load (ETL)/Extract, Load, Transform (ELT)) als visualisatie laag.
    b. Voldoet aan **AVG-richtlijnen en andere wet- en regelgeving**. Vastleggen, beschikbaar stellen, opvragen, delen, uitwisselen en bewerken van data vindt plaats binnen wettelijke kaders en taken van de aanbiedende en/of vragende partijen.

4. **Toegankelijk.**
    a. Toegankelijk en goede mogelijkheden voor **ontsluiting naar buiten** om ervoor te zorgen dat data bevrijd wordt uit applicaties.
    b. Output moet **visueel vormgegeven** kunnen worden (bijv. dashboards).

[3] In tegenstelling tot zogenaamde 'horizontale' data-analyse waarbij algoritmes of technieken worden gebruikt om data van personen te analyseren die in slechts één bron voorkomen (bijv. Federated Learning: training van een algoritme op data van verschillende ziekenhuizen, waarbij patiënten slechts in één ziekenhuis aanwezig zijn)

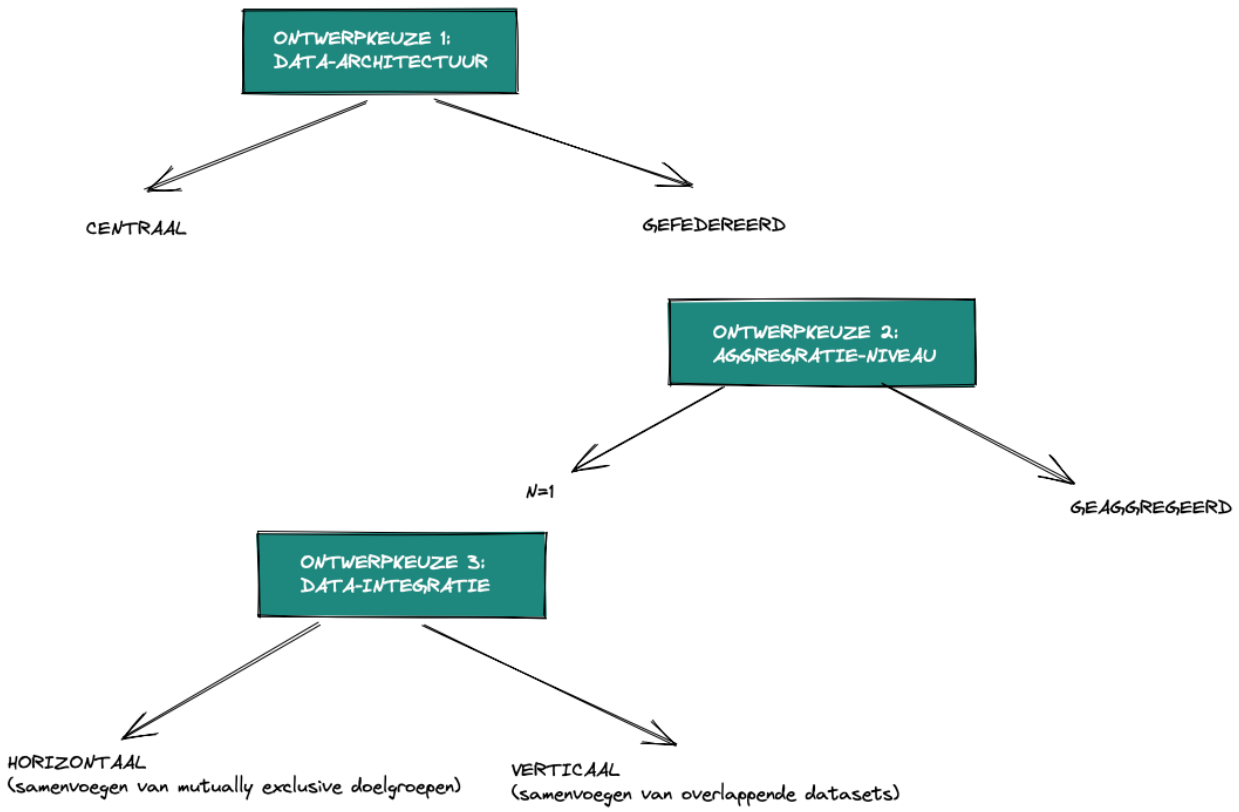
2 Geïntegreerde regionale data-infrastructuur

Het inrichten van een (regionale) data-infrastructuur is maatwerk en kan verschillen per regio. Tegelijkertijd zijn er ontwerpkeuzes die kunnen zorgen voor een gemeenschappelijk fundament, ongeacht de regio. Doel- en informatievragen vertalen zich via specificaties naar een gewenst ontwerp voor de data-infrastructuur.

Hierbij zijn naar ons idee in ieder geval drie fundamentele ontwerpkeuzes te maken:

1. Data-architectuur (inclusief opslag van data): centraal versus gefedereerd
2. Aggregatieniveau (detailniveau onderzoek & output): n = 1 versus geaggregeerd
3. Data-integratie: horizontaal versus verticaal

Bovenstaande drie ontwerpkeuzes dienen als leidraad voor de opbouw van dit stuk en de technieken die hierin wel/niet besproken worden. De ontwerpkeuzes hebben we uitgewerkt in een diagram, waarbij ontwerpkeuze 2 en 3 gebonden zijn aan een bovenliggende keuzemogelijkheid (Figuur 2). Deze keuzes zijn niet per se afhankelijk van elkaar, maar leiden in onze ogen wel tot een optimaal ontwerp.



Figuur 2: Overzicht van de drie ontwerpkeuzes die als kapstok dienen in deze whitepaper. Elke ontwerpkeuze wordt gevolgd door twee mogelijkheden. Er is te zien dat ontwerpkeuze 2 en 3 gebonden zijn aan een bovenliggende keuzemogelijkheid die in onze ogen leidt tot een optimaal ontwerp

Voor iedere (combinatie van) ontwerpkeuze(s) worden in deze whitepaper bekende en beschikbare oplossingen in kaart gebracht. Al deze oplossingen worden beschreven in termen van en/of getoetst op de aspecten zoals weergegeven in de paragrafen hieronder, met inachtneming van de specificaties voor GERDA uit paragraaf 1.2.

2.1 Achtergrond

- Ontstaan (historie, wettelijke grondslag).
- Procedures.
- Volwassenheid: kinderschoenen vs. bewezen?
- Aggregatieniveau input en output.
- Bekende toepassingen (voorbeelden).

2.2 Data governance

- Beheer en inrichting (projectmatig versus duurzaam).
- Eigenaarschap/rol van de deelnemende organisaties.
- Vereisten aan te leveren data.
- Vereisten aan onderzoeksopzet.
- Vereisten aan uitvoeren onderzoek.

2.3 Technologie

- Opslagvorm (centraal versus gefedereerd).
- Mogelijkheden te analyseren data (getallen, tekst, beeld).
- Data-opslag/model (flat file, Pandas, SQL, etc).
- Mogelijkheid gebruik van data science tools (Python, R, etc).
- Output - en publicatiemogelijkheden.

2.4 Beveiliging en juridische/privacy aspecten

- Welk wettelijk kader is de basis voor dit model?
- Omvang van de beveiliging (inhoud van de data, Inhoud van de vraag zoals script/algoritme, ongewenste onthulling van het antwoord en communicatie tussen partijen).
- Werking van beveiliging (standaarden, protocollen).
- Wie is er verantwoordelijk voor de beveiliging (partijen zelf of andere partijen?).

2.5 Ethische en sociale aspecten

- In hoeverre is de structuur: Fair, Accurate, Confidential, Transparent.
- Hanteren van onderzoeksprotocol/procedures.
- Reproduceerbaarheid van modellen/resultaten.
- Ongewenste onthulling door het antwoord?
- Kan je er beleid op voeren welke data je wel/niet kan (laten) analyseren?

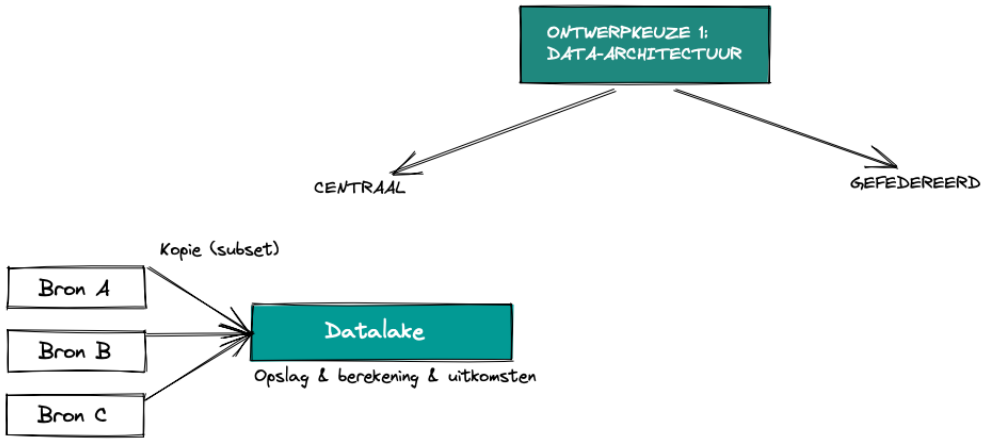
2.6 Relevantie toepassing voor regio’s

- Hoe geeft de oplossing invulling aan de drie - voor onze context - fundamentele ontwerpkeuzes; data governance, aggregatieniveau, data-integratie?

3 Ontwerpkeuze 1: Data-architectuur

Een eerste keuze voor het maken voor een geïntegreerde (regionale) data-infrastructuur is de data regio. Dit begint met de vraag of de data-architectuur centraal of als onderdeel van een gefedereerd stelsel wordt ingericht. Bij centrale data-architectuur wordt (een gedeeltelijke kopie van) de data van deelnemende partijen bij elkaar gebracht op één gedeelde plek. Bij een gefedereerd stelsel blijft de data van iedere organisatie opgeslagen binnen de eigen organisatie en wordt deze 'uitgevraagd' wanneer dat nodig is.

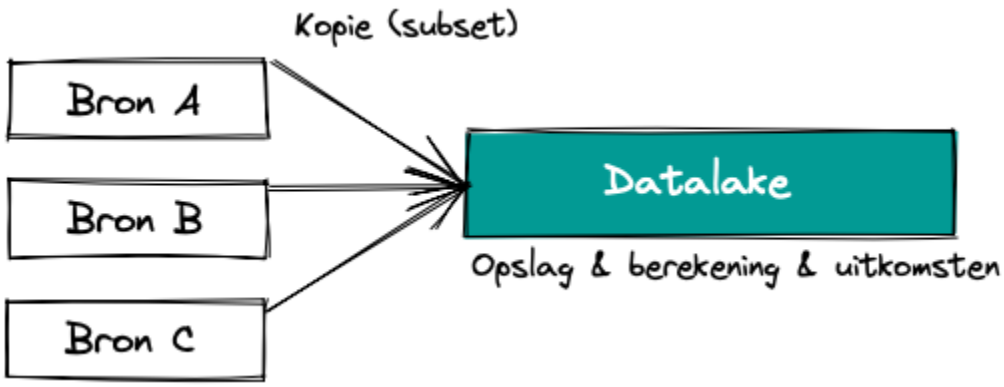
Voor de centrale opslag van data kan een 'datalake' ingericht worden. Dit is een opslagoplossing die wordt gebruikt om grote hoeveelheden ruwe data in de oorspronkelijke structuur te verzamelen. Een datalake kan gestructureerde en ongestructureerde data bevatten. Het datalake dient daarnaast als plek voor domeinoverstijgende analyse en onderzoek op de gecombineerde data van de deelnemende organisaties, maar valt altijd onder centrale data-architectuur (zie ook Figuur 3). Het inrichten van een datalake kan op verschillende niveaus, variërend van nationaal tot regionaal niveau. In paragraaf 3.1 volgt een uitwerking van de (on)mogelijkheden van een nationaal datalake (CBS Remote Access (RA)/microdata-omgeving) en in paragraaf 3.2 wordt de uitwerking beschreven voor een regionaal datalake (Transmuraal Netwerk Midden-Holland).



Figuur 3: Schematisch overzicht van ontwerpkeuze 1 over data-architectuur, inclusief een uitwerking voor centrale data-architectuur. Deze schematische weergave van een datalake laat de relatie zien tussen een drietal bronnen (bron A, B en C) en het datalake. Er is sprake van een éénrichtingsverkeer vanuit de bronnen waarbij een kopie (ook wel subset) op het datalake wordt opgeslagen. Hier vindt opslag plaats en kunnen berekeningen worden gedaan en uitkomsten worden geanalyseerd.

3.1 Nationaal Datalake: CBS-RA omgeving

Achtergrond
Een datalake is een opslagoplossing die wordt gebruikt om grote hoeveelheden ruwe data in de oorspronkelijke structuur te verzamelen. Een datalake kan gestructureerde en ongestructureerde data bevatten. Een datalake is geen vervangend opslagsysteem voor de brondata, maar dient als een plek waar data bijeen wordt gebracht voor analyse en onderzoek (zie Figuur 4). Door data van diverse organisaties bij elkaar te brengen in een centraal datalake ontstaat de mogelijkheid om domeinoverstijgend onderzoek te doen op deze gecombineerde data. Dit kan op verschillende niveaus, variërend van nationaal tot lokaal niveau.

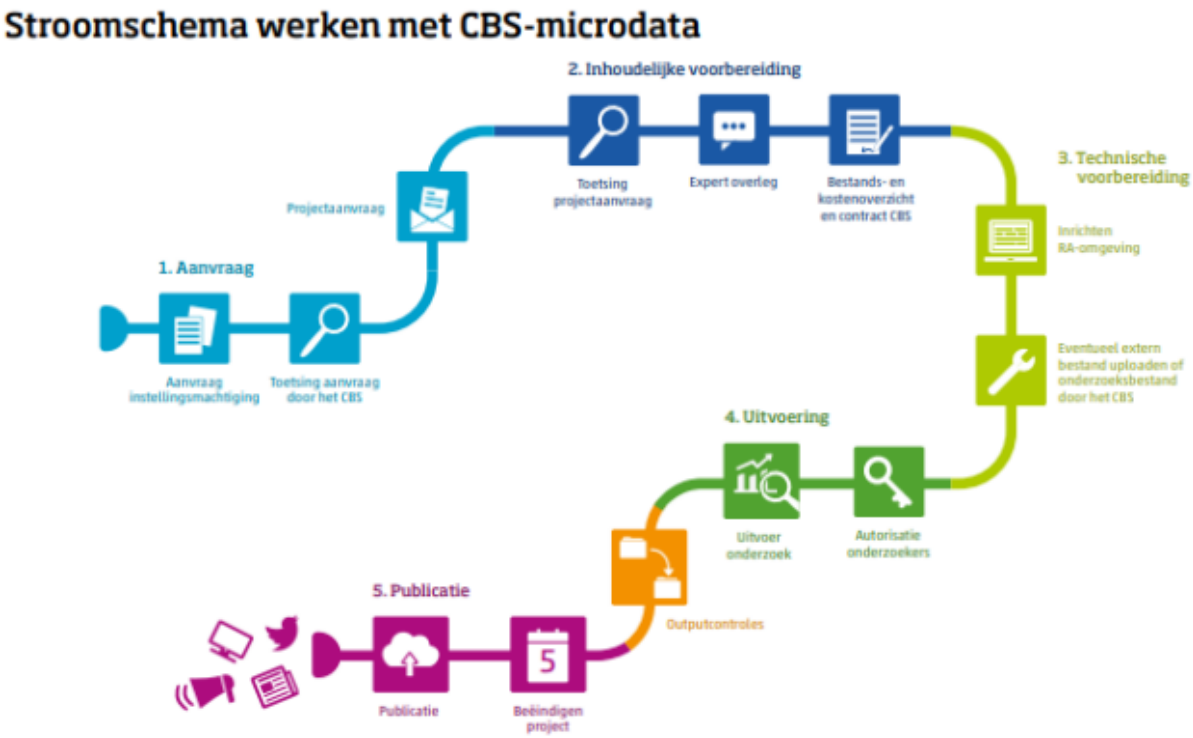


Figuur 4: Deze schematische weergave van een datalake laat de relatie zien tussen een drietal bronnen (bron A, B en C) en het datalake. Er is sprake van een éénrichtingsverkeer vanuit de bronnen waarbij een kopie (ook wel subset) op het datalake wordt opgeslagen. Hier vindt opslag plaats en kunnen berekeningen worden gedaan en uitkomsten worden geanalyseerd.

Er zijn verschillende voorbeelden van een centraal datalake op nationaal niveau, waarvan de meeste voorbeelden van overheidsinstellingen zijn. Er zijn echter maar een beperkt aantal omgevingen waar derden, zoals onderzoekers, (onder strikte voorwaarden) toegang toe hebben. Het bekendste voorbeeld van een nationaal datalake, dat ook gebruikt kan worden voor domeinoverstijgend gezondheidsonderzoek, is de omgeving van het Centraal Bureau voor de Statistiek (CBS): CBS Remote Access (CBS-RA) of CBS microdata-omgeving.

Het CBS heeft als wettelijke taken het van overheidswege (1) verrichten van statistisch onderzoek ten behoeve van praktijk, beleid en wetenschap en het openbaar maken van deze statistieken en (2) het bevorderen van statistische informatievoorziening. Sinds een aantal jaar kan onderzoek worden gedaan met microdata van het CBS. Microdata is koppelbare data op persoons-, bedrijfs- en adresniveau die gebruikt mogen worden voor statistische doeleinden. Nederlandse universiteiten, wetenschappelijke organisaties en planbureaus kunnen hier onder strikte voorwaarden gebruik van maken.

Het werken met CBS data in een CBS microdata-omgeving vraagt om een secure gang van zaken. In Figuur 5 hieronder zijn de stappen die daarvoor moeten worden doorlopen weergegeven.



Figuur 5: Stroomschema van het proces voor het werken met CBS-microdata. In dit stroomschema zijn een aantal fasen beschreven: 1) Aanvraag, 2) Inhoudelijke voorbereiding, 3) Technische voorbereiding, 4) Uitvoering en 5) Publicatie.

Data governance

De CBS microdata-omgeving is een voorbeeld van een nationaal datalake, waarin gewerkt wordt volgens een projectmatige structuur. Er is voor deze omgeving sprake van een hoge mate van centrale governance, uitgevoerd op gestructureerde wijze door de beheerorganisatie van het CBS. Er is een stroomschema met alle stappen die doorlopen worden in het werken met de CBS microdata-omgeving en voor iedere stap zijn heldere afspraken, formulieren, protocollen, richtlijnen en maatregelen beschreven. Waar CBS de verantwoordelijkheid draagt voor de kwaliteit van de CBS-microdata is het projectteam zelf verantwoordelijk voor de kwaliteit van de gebruikte eigen en/of externe databestanden. Voor het indienen van een onderzoek bij CBS moet een instelling geaccrediteerd zijn. Dit geldt voor de meeste kennisinstellingen, zoals universiteiten, maar ook derde partijen kunnen een (onderzoeks)accreditatie aanvragen. De verplichte handmatige outputcontrole en output eisen maken de omgeving minder geschikt voor dynamische, complexe en selectieve vragen, voor sommige visualisatie doeleinden of voor data teruglevering aan het primaire proces op n = 1 niveau.

Technologie

Als een organisatie beschikt over een geldige machtiging voor toegang tot de microdata-omgeving van CBS kan een projectaanvraag gedaan worden, met een onderzoeksomschrijving. Na goedkeuring worden binnen de CBS microdata-omgeving de aangevraagde CBS microdatabestanden (op n=1 niveau) beschikbaar gesteld. Daarnaast kunnen geselecteerde eigen en/of externe databestanden in de CBS microdata-omgeving geïmporteerd en aan de CBS microdata gekoppeld worden. Na uitvoering van het onderzoek mag (na outputcontrole door CBS) geaggregeerde, niet tot personen herleidbare, output worden geëxporteerd. Openbare (wetenschappelijke) publicatie van de uitkomsten van het onderzoek is verplicht.

Voor ieder project vindt centrale opslag van data plaats binnen de CBS microdata-omgeving. In deze projectomgeving staan de benodigde CBS microdatabestanden, met daaraan gekoppeld eventuele geïmporteerde eigen en/of externe bestanden. Bij wijzigingen van de eigen/externe bestanden moeten deze opnieuw geïmporteerd worden. De projectomgeving bevat een verzameling van gestructureerde numerieke en/of tekstuele databestanden en heeft geen onderliggend overkoepelend datamodel. De projectomgeving biedt standaard verschillende (klassieke) analyse-tools, waaronder Excel, SPSS, Stata en R. Op aanvraag kunnen ook softwarepakketten als MLWin worden toegevoegd. Output is gebonden aan technische richtlijnen op het gebied van grootte (<5 mb), formaat (alleen .xls, .doc, .txt, .csv, .bmp, .jpg, .gif) en een aggregatieniveau van n>10. Er vindt een verplichte handmatige outputcontrole plaats door CBS voor het vrijgeven van de uitkomsten. Na afloop van het project wordt toegang tot de projectomgeving afgesloten. Een projectarchief wordt 5 jaar bewaard. Door de centrale opzet (onder beheer CBS) en veiligheidseisen is er geen simpele manier voorhanden waarop data van externe en/of eigen bronnen continu ingelezen kan worden (bijv. door API-koppelingen). Een centraal datalake zoals CBS is daardoor bruikbaarder voor langer durende (onderzoeks) trajecten waarbij de tijdigheid van gegevens minder van belang is en minder geschikt voor doorlopende informatievragen (bijv. rapportages op wekelijks of maandelijks niveau).

Beveiliging en juridische/privacy aspecten

Toegang tot de CBS microdata-omgeving voldoet aan de huidige veiligheidseisen[4]. Alleen aan het onderzoek gekoppelde onderzoekers kunnen, met een persoonsgebonden token, op de CBS microdata-omgeving in loggen via een beveiligde internetverbinding. Onderzoekers krijgen hierbij toegang tot alleen de voor het onderzoek noodzakelijke bestanden en alle microdata blijft binnen de beveiligde omgeving van het CBS. Om dit te garanderen vindt outputcontrole plaats door het CBS. Ook zijn alle onderzoekers gebonden aan gebruikersregels en vindt er logging plaats van alle handelingen van iedere RA microdata-onderzoeker. Bij overtreding heeft het CBS een maatregelenbeleid.

Toegang tot de CBS microdata-omgeving voldoet aan de huidige privacy-eisen. De omgeving valt onder de wet op de statistiek, waardoor geen informed consent nodig is. Hier staat tegenover dat ieder onderzoeksvoorstel vooraf wordt getoetst door CBS (en bij academische onderzoeken ook vaak nog door een medisch-ethische commissie). Ook wordt een juridische overeenkomst getekend voor het

[4] Eindrapport naar onderzoek 'Remote Acces to CBS-microdata'. Te downloaden van CBS microdata-website.

verkrijgen van toegang tot de omgeving. Tot slot staan het borgen van de privacy en het voorkomen van onthulling van personen of bedrijven centraal in de inrichting van het gehele proces rondom het gebruik van microdata.

Ethische en sociale aspecten

Governance op de CBS microdata-omgeving is zodanig ingericht dat (grotendeels) wordt voldaan aan de FACT uitgangsprincipes[5]:

- **F**air/eerlijk; goedgekeurde onderzoeksopzet nodig, inclusief toestemming Medisch-Ethische Commissie (MEC).
- **A**ccurate/nauwkeurig; richtlijnen voor het minimum aantal records in een groep, het minimale aandeel van een groep ten opzichte van de gehele onderzoekspopulatie en voor een minimum aantal vrijheidsgraden in een (regressie)model.
- **C**onfidential/vertrouwelijk; er vindt, voordat gegevens de CBS microdata-omgeving verlaten, outputcontrole plaats op herleidbaarheid tot een individu.
- **T**ransparant; hier is geen standaard voorziening voor vanuit CBS. De verantwoordelijkheid ligt bij projectteam (o.a. controleerbaarheid syntax, reproduceerbaarheid van de uitkomsten).

Relevantie geïntegreerde regionale data-infrastructuur

Minder relevant.

Data-architectuur

Technisch gezien is een centraal datalake een geschikte en relevante optie als opslagvorm voor een geïntegreerde regionale data-infrastructuur. Hiervoor is het echter wel nodig dat (handmatig) veel data verplaatst wordt.

Aggregatieniveau

Volgens de CBS-richtlijnen is het niet mogelijk om te exporteren naar een aggregatieniveau lager dan n=10.

Data-integratie

Technisch gezien is het mogelijk om zowel horizontaal als verticaal onderzoek toe te passen.

[5] https://www.thedigitalsociety.info/nl/over/dataprincipes/

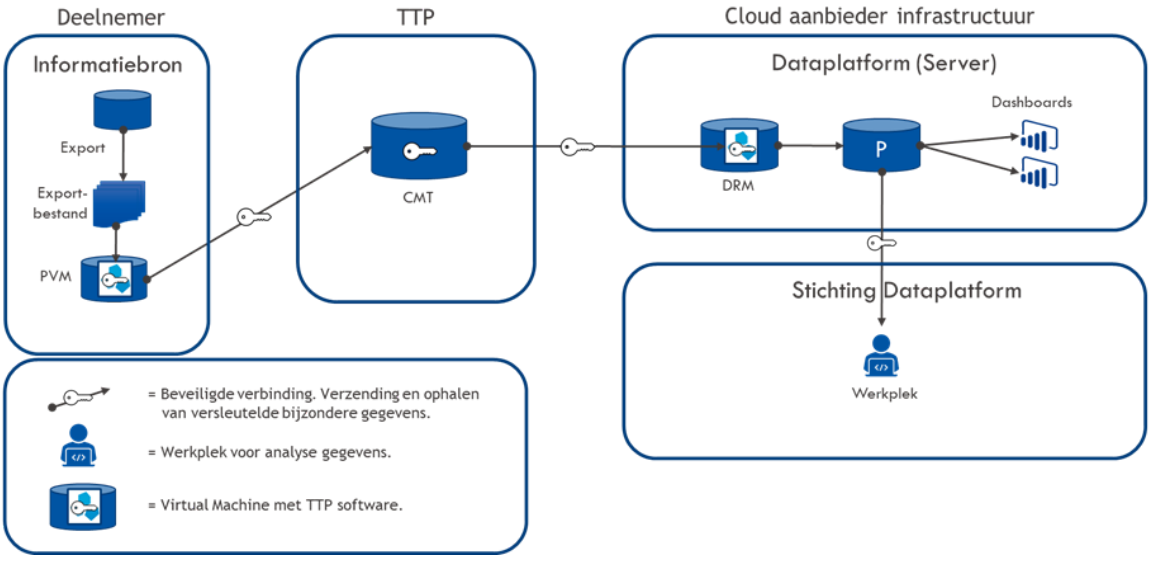
3.2 Regionaal Datalake: Transmuraal Netwerk

Achtergrond

Een datalake kan ook op regionaal niveau ingericht worden. Een voorbeeld hiervan vinden we in de regio Midden-Holland binnen het Transmuraal Netwerk. Diverse samenwerkende zorgaanbieders en gemeenten in Midden-Holland hebben als gezamenlijke missie regionaal samen te werken aan goede zorg en welzijn. Om deze missie te bereiken is domeinoverstijgende samenwerking nodig. Hier wordt binnen het Transmuraal Netwerk vorm aan gegeven door middel van een gezamenlijk onderzoek (leidend tot nieuwe domeinoverstijgende inzichten) die de besturen van de samenwerkende partijen in staat stellen om datagedreven besluiten te nemen. De verwerkingsgrond voor de gegevens betreft een statistisch onderzoek. Een onderzoeksprotocol is van tevoren opgesteld.

Technologie

Voor het onderzoeksproject van het Transmuraal Netwerk Midden-Holland vindt centrale opslag van data plaats binnen een gezamenlijk regionaal dataplatform. Het ontsluiten van de gegevens uit het bronsysteem van de deelnemende partijen vindt plaats door de deelnemende partijen zelf. Bestanden worden dubbel gepseudonimiseerd (via een trusted third party) en op persoonsniveau aangeleverd (n=1) en met elkaar gekoppeld (zie Figuur 6). Bij wijzigingen van de bestanden moeten deze opnieuw geïmporteerd worden. Resultaten worden aan de deelnemende partijen beschikbaar gesteld via dashboards en rapportages op basis van geaggregeerd niveau (n>10). Na afloop van het onderzoek worden de gegevens vernietigd.



Figuur 6: Een schematische weergave van de datastromen van het Transmuraal Netwerk Midden-Holland. Exportbestanden met relevante gegevens worden door de deelnemende partijen via een Trusted Third Party (TTP) gekopieerd naar een centraal dataplatform. Het dataplatform voorziet in zowel centrale data-architectuur

(datalake) als een datawerkplaats waar de geautoriseerde onderzoekers analyses kunnen doen en visualisaties kunnen maken.

Data governance

Het dataplatform van het Transmuraal Netwerk Midden-Holland is een voorbeeld van een regionaal datalake, waarin gewerkt wordt volgens een projectmatige structuur. Er is voor deze omgeving sprake van een hoge mate van centrale governance, uitgevoerd door een door de deelnemende partijen opgerichte stichting. Deze stichting heeft als opdracht de infrastructuur te beheren die noodzakelijk is voor het dataplatform en om het onderzoek uit te voeren ten behoeve van de deelnemende partijen. Door de stichting worden de diensten van een leverancier ingehuurd voor ondersteuning bij de bouw, inrichting en het beheer van het dataplatform. Afspraken over de aanlevering van de data (kwaliteit, manier van pseudonimisatie, werken met een Trusted Third Party (TTP), etc.) en de gouden standaard voor gebruikte (persoons)gegevens worden vastgelegd door de deelnemende organisaties.

Beveiliging en juridische/privacy aspecten

Vanaf de werkplek van de onderzoeker(s) is er beveiligde toegang tot de verzamelde gegevens op het dataplatform om deze te combineren, ordenen en analyses te verrichten zoals beschreven in het onderzoeksprotocol. Alleen geautoriseerde gebruikers hebben (op basis van IP white listing) toegang tot het data platform. Als het gaat om detachering wordt een geheimhoudingsverklaring getekend door de onderzoeker. Alle onderzoekers worden op de hoogte gesteld van het beveiligingsbeleid dat geldt voor het gebruik van de infrastructuur. De data blijven binnen de gedeelde omgeving voor de duur van het onderzoek en worden niet doorgezonden, verspreid en kunnen niet door derden worden opgevraagd.

Het versturen van de gegevens naar de TTP en van de TTP naar het data platform gebeurt via een beveiligde verbinding. Daarnaast vindt bij het aanleveren naar de TTP en vervolgens door de TTP pseudonimisatie plaats. Omdat de stichting niet zelf over de capaciteit beschikt om de infrastructuur op te zetten en/of te beheren wordt dit uitbesteed aan een leverancier. Verwerkersovereenkomsten zijn afgesloten tussen de deelnemende organisaties, de TTP en de leverancier van het dataplatform. Hierin zijn ook afspraken gemaakt over de beveiligingsnormen (waaronder ISO27001 en NEN7510).

Voor het onderzoek van het Transmuraal Netwerk Midden-Holland is een DPIA uitgevoerd, waarmee voorafgaand aan de verwerking de privacy risico's voor betrokkenen in kaart is gebracht. Deze DPIA wordt gezien als een iteratief proces dat opnieuw wordt uitgevoerd zodra nieuwe dataleveranciers toetreden, afspraken wijzigen of de inhoud van de verzamelde data wijzigt. Voor uitvoering en sturing op het onderzoek zijn toezichthoudende en sturende organen in het leven geroepen om het onderzoek in goede banen te leiden en toe te zien op de naleving van juridische voorwaarden (zoals rechtmatigheid voor het verwerken van persoonsgegevens, recht op inzage, rectificatie, vergetelheid, beperking van de verwerking en overdraagbaarheid, up-to-date verwerkersovereenkomsten, toegangsrechten en andere afspraken/maatregelen).

Ethische en sociale aspecten

Governance op het regionale dataplatform is de verantwoordelijkheid van de deelnemende partijen binnen het Transmuraal Netwerk Midden-Holland. Onderlinge afspraken zijn nodig om aan de FACT uitgangsprincipes (Fair, Accurate, Confidential, Transparant - zie ook voetnoot 4 bij nationaal datalake) te voldoen. Het werken met een onderzoeksopzet en richtlijnen voor het minimum aantal records in een groep (n>10) waarover uitkomsten gepresenteerd worden zijn hier voorbeelden van.

Relevantie geïntegreerde regionale data-infrastructuur

Minder relevant.

Data-architectuur

Een regionaal datalake een geschikte en relevante optie als opslagvorm voor een geïntegreerde regionale data-infrastructuur. Hiervoor is het echter wel nodig dat (handmatig) veel data verplaatst wordt.

Aggregatieniveau

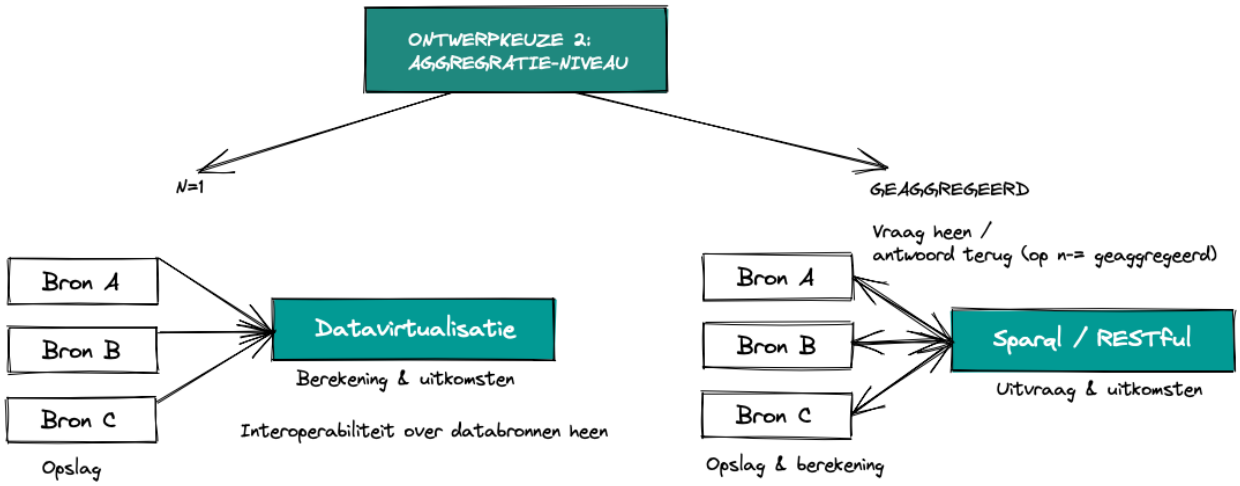
Technisch gezien is het mogelijk om data te verwerken op aggregatieniveau n=1. Het laagste detailniveau wordt uiteraard bepaald door het detailniveau van de brondata. Binnen het Transmuraal Netwerk heeft men de keuze gemaakt om geen gegevens te exporteren met een aggregatieniveau lager dan n=10.

Data-integratie

Het is mogelijk om zowel horizontaal als verticaal onderzoek toe te passen.

4 Ontwerpkeuze 2: Aggregatieniveau

De volgende fundamentele ontwerpkeuze richt zich op het aggregatieniveau van de benodigde gegevens. Welk detailniveau biedt onderzoekers de juiste mogelijkheden voor hun analyses en voorziet onderzoekers en andere gebruikers van zinvolle uitkomsten, om een 'call to action' op te baseren. Voor een beeld van de samenhang tussen de diverse gebruikersgroepen en het gewenste detailniveau van onderzoek zie Figuur 7. De twee hierin gepresenteerde mogelijkheden zijn FAIR data stations en datavirtualisatie. Deze mogelijkheden staan in dit hoofdstuk beschreven.



Figuur 7: Schematisch overzicht van ontwerpkeuze 2 over detailniveau, inclusief een uitwerking voor geaggregeerde data-output. Dit is een schematische weergave van het leveren van geaggregeerde output over een drietal bronnen (bron A, B en C) binnen verschillende organisaties met een sparql/RESTful data station en met datavirtualisatie. Bij sparql/RESTful is er sprake van een tweerichtingsverkeer tussen de bronnen en het station, waarbij het antwoord op de vraag retour gestuurd wordt (niet een dataset). Bij datavirtualisatie is sprake van éénrichtingsverkeer.

Een voorbeeld van een onderzoeksvraag op geaggregeerd niveau is: 'wat is voor doelgroep x het percentage valincidenten over periode y?'. Hiervoor is het laagst gewenste detailniveau de gehele *doelgroep x*. Andere onderzoeksvragen zijn gedetailleerder en betreffen kenmerken op persoonsniveau (gekoppeld over bronnen heen). Een voorbeeld van dit type onderzoeksvraag is: 'wat zijn kenmerken van cliënten die in zorg zijn geweest bij organisatie x en binnen 3 maanden in zorg komen bij organisatie z? '. Een gevolg van onderzoek op een lager detailniveau, betekent data met een hogere kans op identificatie van individuele personen en vereist een andere data governance met (nog) striktere eisen aan databeveiliging.

4.1 FAIR data stations

Achtergrond
Een data station is een infrastructureel component om gefedereerd (een selectie van) gegevens in een bron toegankelijk te maken. Een data station is daarbij vaak onderdeel van een bredere data-infrastructuur, waarbij een data station per organisatie dient als ontsluitingspoort van data, terwijl de data bij de bron blijft staan.

Er zijn verschillende soorten invullingen van data stations variërend van locaties waar simpelweg data toegankelijk wordt gemaakt, tot locaties waar ook complexe berekeningen kunnen worden uitgevoerd. In de eindrapportage van het Zorginstituut Nederland voor Personal Health Train (een afsprakenstelsel voor gefedereerde data-analyse) naar FAIR data stations worden drie verschillende typen data station onderscheiden: sparql-endpoint, RESTful-interface en docker container. Een omschrijving van deze drie typen data station staat hieronder:

FAIR data station *- door Zorginstituut voor Personal Health Train*

Het station dat diensten aanbiedt om bij de data te komen. Bij de implementatie zijn wij tot de conclusie gekomen dat er diverse typen diensten zijn en daarmee meerdere technische implementaties. Afhankelijk van de gebruikstoepassing moet de dienst meer of minder beveiligd zijn en afhankelijk van de complexiteit van de analyse moet de dienst meer of minder mogelijkheden bieden.

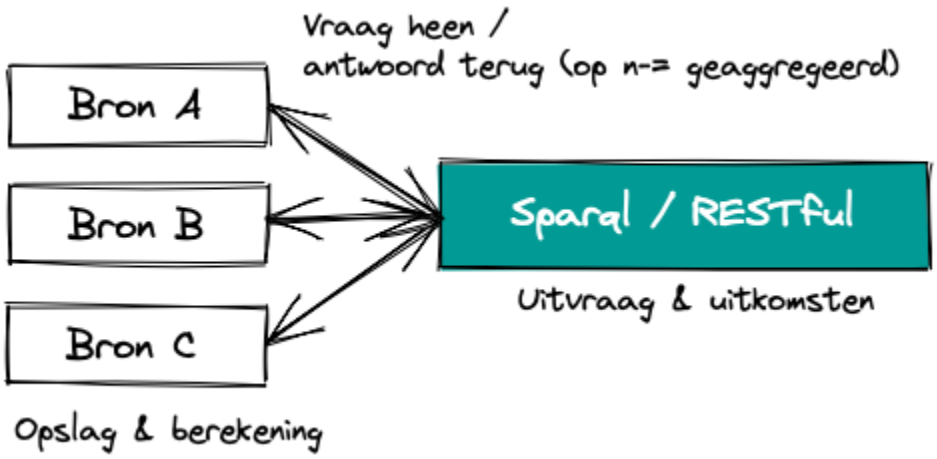
Wij hebben drie type diensten gerealiseerd die een datastation kan aanbieden.

1. **Sparql-endpoint** voor vragen waarbij privacy minder van belang is, zoals bij *open data*, maar waarbij wel veel mogelijkheden worden geboden aan de analist om zelf zijn vragen te definiëren. Deze dienst is relatief eenvoudig om te implementeren en biedt de gelegenheid om nieuwe gebruikstoepassingen voor data te ontwikkelen. In de casus heeft elk station een Sparql-endpoint.

2. **RESTful-interface** om privacy goed te kunnen waarborgen en vragen vooraf te definiëren. Deze dienst biedt weinig flexibiliteit voor de analist, is iets complexer om te implementeren dan een Sparql-endpoint, maar geeft veel zekerheid vanuit beveiliging en privacy. In de casus bieden de stations van de ambulance diensten via deze interface, naast het Sparql-endpoint.

3. **Docker container** In deze omgeving kan de analist een algoritme draaien bij de data-eigenaar. Deze dienst komt het dichtst bij de metafoor van de trein. In de praktijktoets is voor een *docker container* gekozen als locatie van de algoritmes omdat de technologie de software onafhankelijk en autonoom maakt van de omgeving waarin het moet opereren. Er zijn goede waarborgen mogelijk voor privacy én de analist heeft veel mogelijkheden om de analyse vorm te geven. Deze dienst is technisch gezien complexer om te implementeren dan de eerste twee diensten. Ook zijn afspraken nodig tussen de gebruikers van de diensten om de juiste waarborgen te creëren op het gebied van privacy, techniek en organisatie. In de implementatie van de casus bieden de stations van beide ziekenhuizen deze interface.

Het Sparql-endpoint is eenvoudig in implementatie, maar biedt minimale beveiliging. Hierdoor kan het slechts op beperkte soorten van geaggregeerde data worden toegepast. Een RESTful-interface is weliswaar veiliger, maar nog steeds alleen geschikt voor geaggregeerde data. In Figuur 8 zijn deze beide opties opgenomen. De docker container is de best beveiligde vorm van een data station en maakt het veilig uitvoeren van gefedereerde berekeningen op basis van algoritmen op een laag detailniveau (n=1) wel mogelijk. Omdat we ons in dit hoofdstuk richten op uitkomsten op geaggregeerd detailniveau wordt de docker container (als vorm van Federated Learning) verder meegenomen in hoofdstuk 5, waar de oplossingen voor detailniveau n=1 aan bod komen.



Figuur 8. Deze weergave laat de relatie zien tussen een drietal bronnen (bron A, B en C) en een Sparql/RESTful data station. Er is sprake van een tweerichtingsverkeer tussen de bronnen en het station, waarbij het antwoord op de vraag retour gestuurd wordt (niet een dataset).

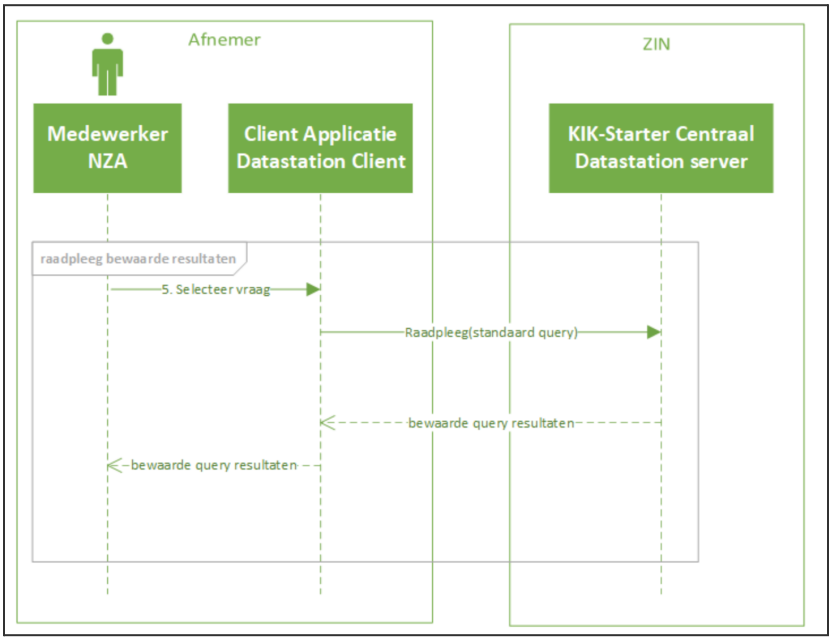
FAIR data stations voor geaggregeerde dataverwerking worden onder andere gebruikt door het programma Keteninformatie Kwaliteit Verpleeghuiszorg (KIK-V). Binnen Personal Health Train is een toepassing neergezet voor een use case van KIK-V, waarbij onderzoeksinstanties in staat worden gesteld om via een fair data station gevalideerde onderzoeksvragen te stellen aan de Verpleeg - en Verzorgingshuizen en Thuiszorg (VVT).

Technologie

Data stations bieden een oplossing om data te analyseren, terwijl deze bij de bron blijft. Er zijn 'off-the-shelve' oplossingen. Belangrijkste constatering is dat toepassing beperkt is doordat berekening bij de bron alleen zinvol is bij datasets van cliënten die elkaar niet overlappen (en dus niet domeinoverstijgend/verticaal) omdat er enkel uitkomsten op geaggregeerd niveau teruggegeven worden. Wel is het een gefedereerde oplossing waarbij de data bij de bron blijft.

Vanuit KIK-V ziet de meest geïntegreerde oplossing om data te delen via FAIR data stations er uit zoals weergegeven in Figuur 9 (variant Sparql endpoint). Bij zowel de afnemer als de aanbieder van data,

respectievelijk de 'client' en de 'server', staat een data station. De medewerker stuurt een vraag uit vanaf de client en krijgt een antwoord op geaggregeerd niveau terug vanuit de server.



Figuur 9. Visuele weergave van infrastructurele interactie tussen afnemer (datagebruiker, hier: Nederlandse Zorgautoriteit) en Zorginstituut Nederland (data verstrekker).

Data governance

Bij FAIR data stations wordt de data governance in een initiële data verwerkersovereenkomst geborgd. Na de initiële inrichting van de FAIR data stations zal de datavrager per uitvraag een zogenaamd attest moeten 'overhandigen' om aan te tonen dat de uitvraag past binnen de afspraken en is gevalideerd.

Beveiliging en juridische/privacy aspecten

De mate van beveiliging van FAIR data stations varieert afhankelijk van het type station (en de mogelijkheden die daarbij horen). Een Sparql-endpoint is zwak beveiligd, maar ondersteunt ook alleen beperkte vragen op een hoog aggregatieniveau. Een RESTful-interface biedt meer beveiliging, maar voorziet ook enkel in geaggregeerde output. Een docker container biedt veel beveiliging, waarbij veilig het algoritme heen en op laag aggregatieniveau de data retour gestuurd kan worden.

Personal Health Train geeft invulling aan 'privacy by design' door data niet te verplaatsen en daarnaast berekeningen bij de bron te laten plaats vinden.

Ethische en sociale aspecten

Governance in de FAIR data stations is zodanig ingericht dat als volgt wordt voldaan aan de FACT uitgangsprincipes[6]:

- **F**air/eerlijk; een datavraag moet worden vergezeld door een attest om aan te tonen dat de datavrager een valide vraag stelt.
- **A**ccurate/nauwkeurig; de afsprakenset KIK-V definieert welke definitie van de gegevens en welke onderzoeksparameters er van toepassing op de datavraag zijn (denk o.a. aan peildata, scope).
- **C**onfidential/vertrouwelijk; In het hierboven genoemde attest wordt niet alleen een valide datavraag gecontroleerd, maar ook of de datavrager gelegitimeerd een datavraag mag stellen.
- **T**ransparant; de ontvangende organisatie van een datavraag kan de datavraag beoordelen op juistheid. Door de mate van automatisering is reproduceerbaarheid van uitkomsten en logging geborgd.

Relevantie geïntegreerde regionale data-infrastructuur

Relevant, maar beperkt toepasbaar.

Data-architectuur

Data architectuur vindt plaats aan de bron en is dus gefedereerd.

Aggregatieniveau

Een Sparql endpoint en RESTful interface zijn beiden enkel op geaggregeerd niveau toepasbaar. Een docker container kan wel op individueel niveau inzichten geven.

Data-integratie

Belangrijkste bezwaar vanuit regionale data-infrastructuur is de beperkte toepassing door gefedereerd berekeningen uit te laten voeren op verticaal niveau, waardoor er een beperking in mogelijkheden voor de regionale toepassing lijkt te zijn.

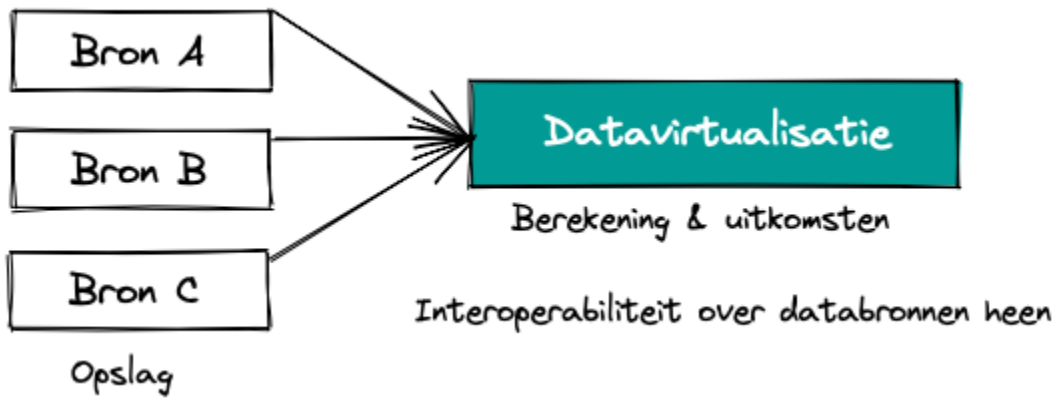
[6] https://www.thedigitalsociety.info/nl/over/dataprincipes/

4.2 Datavirtualisatie

Achtergrond

Datavirtualisatie, of gegevens virtualisatie, is het tijdelijk (voor de duur van de analyse) in een virtuele laag samenbrengen van gegevens vanuit meerdere databronnen. Het is een technische oplossing die is ontstaan in de jaren zestig en nu circa vijftien jaar gebruikt wordt als stabiele technologie in productie omgevingen. Datavirtualisatie is als stand-alone oplossing te realiseren, maar ook als een onderdeel van een bredere oplossing.

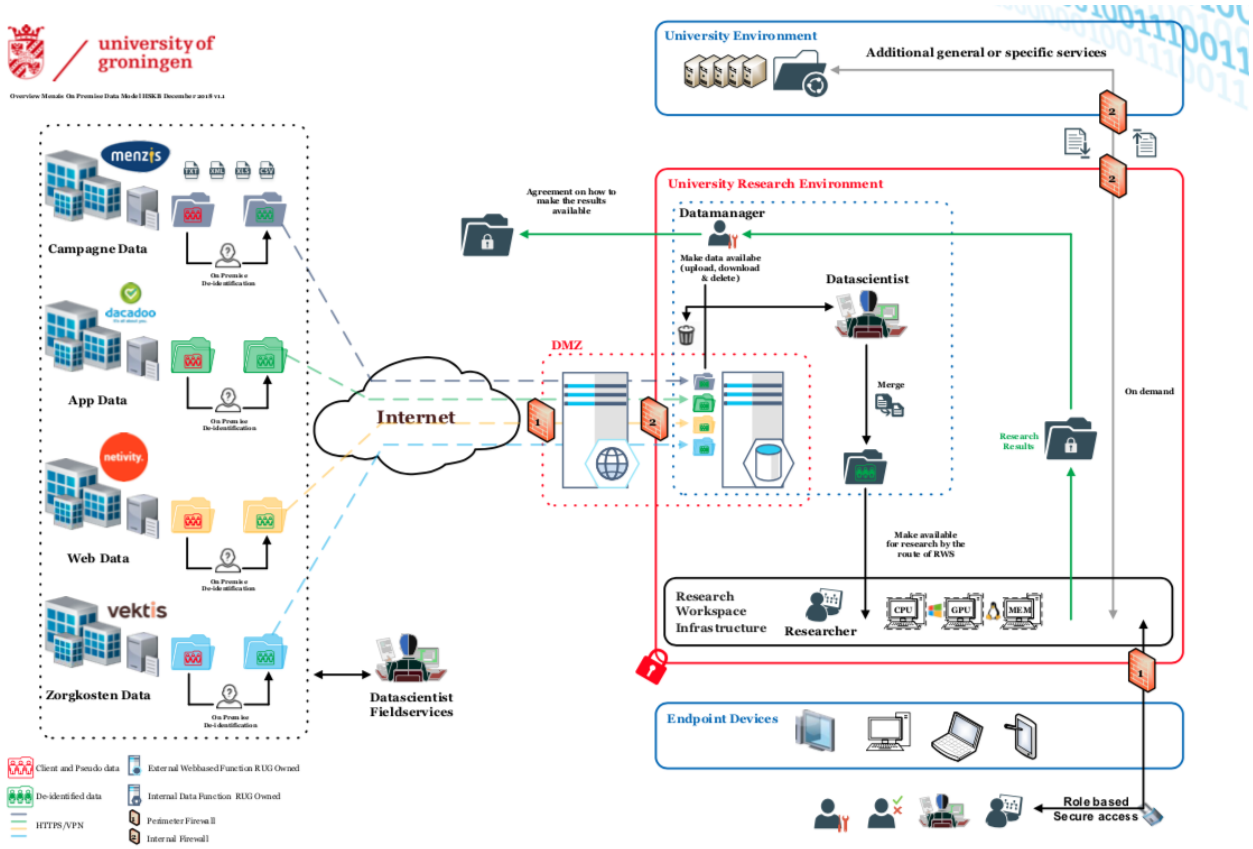
In tegenstelling tot meer traditionele opslagmethodes, zoals bijvoorbeeld een centraal datalake sla je bij datavirtualisatie geen data op, anders dan bij de bron. Waar hardware en software het toelaten kan je met datavirtualisatie wel extraheren en transformeren (de E en de T van ETL), maar blijft data altijd bij de bron en sla je geen kopie van de data op (de L van load). Een tweede voordeel is dat door gebruik te maken van datavirtualisatie meerdere bronnen samen komen in één virtuele laag, waardoor gebruikers van de data maar één ingang nodig hebben om de databronnen aan te spreken. Tot slot is configuratie en het gebruik van front-end BI-applicaties (o.a. rapportage- en dashboard tools) eenvoudiger. Dit maakt dat datavirtualisatie steeds breder gebruikt wordt. Grafisch kan datavirtualisatie worden weergegeven als in Figuur 10.



Figuur 10. Deze weergave laat de relatie zien tussen een drietal bronnen (Bron A, B en C) die gekoppeld zijn aan een virtuele laag (datavirtualisatie). In tegenstelling tot een datalake wordt er geen data opgeslagen, maar blijft de data bij de bron. Een gebruiker kan via de virtuele laag onderzoekshandelingen uitvoeren zoals extractie en transformatie (visueel rechts van het groene kader in bovenstaande figuur).

Technologie

Een bestaande toepassing waarin datavirtualisatie gebruikt wordt is een onderzoeksomgeving binnen de het Centrum voor Informatie Technologie van de Rijksuniversiteit Groningen (RUG-CIT), ingericht door Hans Brouwers (zie Figuur 11). In deze toepassing is de CBS micro-omgeving (onder web data in Figuur 11) één van de databronnen die virtueel beschikbaar wordt gesteld binnen de lokale onderzoeksomgeving van de data scientist. Aangezien de CBS-RA omgeving de meest strikte voorwaarden kent om data veilig te 'ontvangen' en bewerken, ligt in onderstaande beschrijving het accent op de aspecten rondom de hiervoor gekozen inrichting.



Figuur 11. Visualisatie o.b.v. datavirtualisatie van bronnen en invulling datawerkplaats onderzoeksomgeving RUG. Aan de linkerzijde zijn inkomende databronnen getoond, die virtueel samen worden gebracht in een onderzoeksomgeving van de universiteit (rechterzijde van de figuur). Er vindt nadrukkelijk rollen-scheiding plaats in een beveiligde omgeving, waarbij de data scientist toegang heeft tot de brondata en de onderzoeker 'slechts' indirect toegang geeft tot gecombineerde (en mogelijk ook genormaliseerde) data.

Voor de inrichting van de onderzoeksomgeving binnen de RUG lagen zijn de volgende ontwerpprincipes gebruikt voor het technisch ontwerp:

- Beveiliging:
    - Van het ingaande en uitgaande dataverkeer.
    - De werkplek van de onderzoeker.
    - Role based access om verschillende rollen en bijbehorende rechten invulling te geven.

        - Logging.
- Functionaliteiten (voor onderzoek en de onderzoeker):
    - Gelaagde inrichting om data te delen (persoonlijk, groep, 1-op-n en 1-op-1).
    - Toegang vanaf elke geografische locatie.
    - Schaalbare performance en opslag.
    - Selfservice user management.
    - Ondersteuning van meerdere versies van dezelfde applicatie.
    - Werken op meerdere schermen.

Om de benodigde beveiliging te realiseren zijn er diverse infrastructurele zones ingericht binnen de onderzoeksomgeving, zodat bij penetratie van buitenaf het risico op een datalek beperkt is. Het gaat om vier zones, waarbij de inrichting varieert naar BIV-classificatie; untrusted, semi-trusted, trusted en very trusted. Tussen deze verschillende zones zijn diverse functies ingericht als load balancing, filtering, authenticatie en logging.

Data governance

Door gebruik te maken van datavirtualisatie wordt toegang tot (een selectie van) gegevens uit diverse databronnen mogelijk op één plaats in de data architectuur. Deze virtuele laag, die toegang biedt tot alle domeinoverstijgende data, maakt het mogelijk om per databron op het niveau van datavelden toegang te verlenen aan specifieke gebruikersgroepen door rechten per groep toe te kennen. Hierbij is het mogelijk om deze rechten per groep, per databron en per dataveld te variëren. Afstemming over recht tot inzage en bewerking van de data wordt via afspraken bij de inrichting van de virtuele laag ingevuld. Door gebruik te maken van data masking, anonimiseren en pseudonimiseren van gegevens is het daarnaast mogelijk om op één virtuele gegevensset meerdere gebruikersgroepen te autoriseren[7]. Zowel BI-oplossingen als data science-toepassingen kunnen gebruik maken van de virtuele laag.

Beveiliging en juridische/privacy aspecten

Datavirtualisatie biedt zekerheid aan de eigenaren van databronnen dat hun data infrastructureel op veilige wijze wordt samengevoegd. Primair voorziet datavirtualisatie niet in anonimisatie of versleuteling van gegevens.

Bij datavirtualisatie ligt het accent op de beveiliging van de koppelingen naar de databronnen. De beveiliging wordt voornamelijk aangebracht op infrastructureel niveau door zones aan te brengen waarbinnen een dreiging van buitenaf kan worden geïsoleerd. Ook is er een scheiding tussen de virtuele laag en de verschillende (applicaties van) gebruikersgroepen van belang om te voorkomen dat een inbreuk in de beveiliging verstrekkende impact kan hebben. Er vindt geen specifieke beveiliging op dataniveau plaats door bijvoorbeeld pseudonimisatie of versleuteling.

Afstemming over recht tot inzage en bewerking van de data wordt via afspraken bij inrichting van de virtuele laag ingevuld. Het inrichten van specifieke rechten per doelgroep zorgt ervoor dat

[7] Blog: "Met datavirtualisatie maken we zorginformatie sneller en makkelijker beschikbaar" | Vektis.nl

data-gebruikers alleen die data kunnen bekijken en bewerken waar zij toegang toe hebben op basis van hun rol. Deze inrichting op basis van een need-to-know uitgangspunt beperkt het risico van ongeoorloofd datagebruik.

Ethische en sociale aspecten
Datavirtualisatie is op zichzelf geen oplossing om invulling te geven aan (het navolging geven van) onderzoeksprotocollen. Operationalisatie van onderzoeksprotocollen moet primair plaatsvinden door het maken van afspraken rondom data governance. Het gevolg daarvan is dat zelf invulling gegeven moet worden aan aspecten zoals infrastructurele beveiliging van lijnen en de werkplek van de onderzoeker, user management, validiteit van onderzoeksvragen, analyse- en outputniveau van de data en logging. Onderzoeksprotocollen zullen regionaal opgesteld, ingevoerd en gemonitord moeten worden.

Relevantie geïntegreerde regionale data-infrastructuur
Zeer relevant, echter als stand alone biedt het onvoldoende beveiliging van (zorg)data.

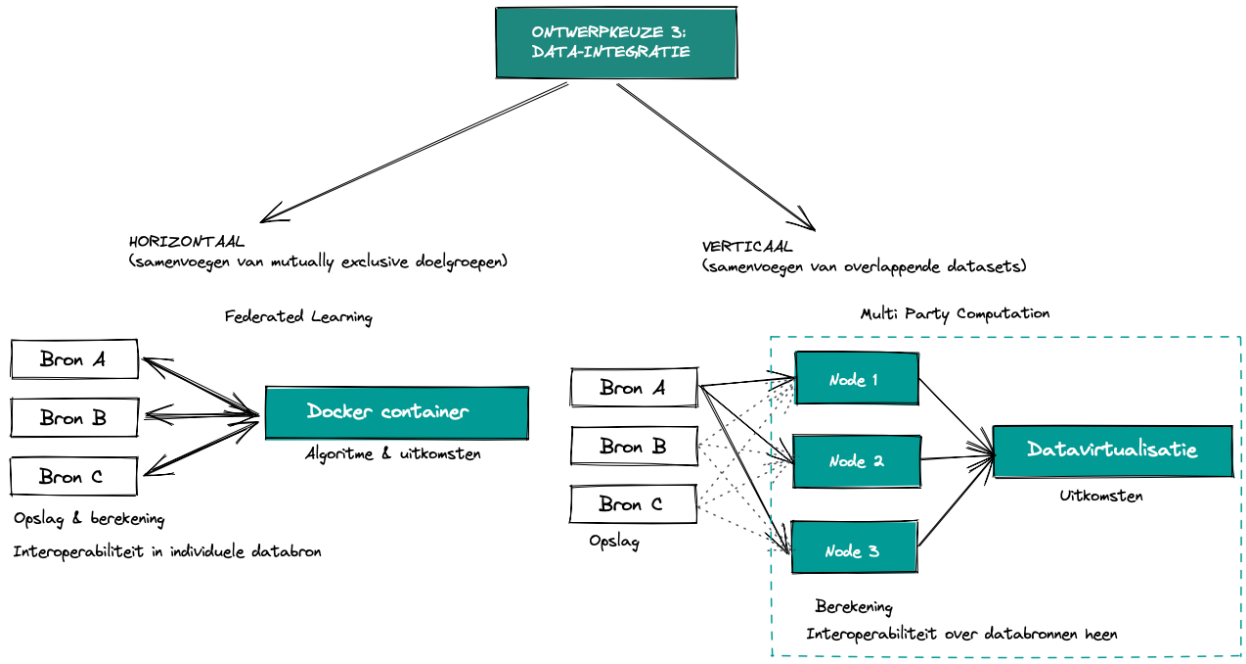
Data-architectuur
Datavirtualisatie is zeer relevant en biedt technisch de mogelijkheden om gefedereerde data te gebruiken. Belangrijk aspect daarbij is dat er centraal geen data wordt opgeslagen.

Aggregatieniveau
Per onderzoeksvraag wordt real time de data uit de bronnen opgehaald en samengebracht op individueel niveau.

Data-integratie
Datavirtualisatie voorziet in horizontale en verticale data-integratie. Het is een maatwerk-oplossing die technisch en qua onderzoeksprotocollen zelf gecreëerd en beheerd moet worden.

5 Ontwerpkeuze 3: Data-integratie

De derde ontwerpkeuze kijkt naar de wijze waarop data geïntegreerd moet worden. We kijken hierbij specifiek naar methodes die dit doen op basis van privacy by design-technologie, waarbij waarborgen van de privacy van personen het uitgangspunt is. Dit perspectief wordt ook wel Privacy Preservering Analytics (PPA) genoemd. We maken hierbij het onderscheid tussen horizontale data-integratie zoals gebruikt bij Federated Learning en verticale data-integratie op basis van Multi Party Computation (MPC). Deze derde ontwerpkeuze is afgebeeld in Figuur 12 .



Figuur 12. Schematisch overzicht van ontwerpkeuze 3 over data-integratie, inclusief een uitwerking voor Federated Learning en Privacy Preserving Analytics.

De eerste vorm van data-integratie is horizontaal. Horizontale data-integratie is het samenvoegen van vergelijkbare data van unieke personen in verschillende datasets. Een praktisch voorbeeld is het bijeen brengen van data van patiënten uit verschillende ziekenhuizen om een algoritme te te trainen voor het opsporen van borstkanker (Federated Learning). In iedere dataset van ieder ziekenhuis (bijvoorbeeld in Rome, Amsterdam of Oslo) zijn de patiënten exclusief, er zijn dus geen patiënten die in meerdere datasets voorkomen. Praktisch ontstaat er door het samenvoegen dus één grote 'lijst' met patiënten die allemaal specifieke kenmerken hebben, afkomstig uit één van de datasets. De integratie beperkt zich tot het samenvoegen van de gehele datasets tot 'één grote dataset'.

Bij verticale data-integratie is er sprake van integratie van gegevens van specifieke personen over meerdere datasets. Er zijn specifieke kenmerken van een persoon die bijvoorbeeld aanwezig zijn in de dataset van het ziekenhuis, terwijl andere kenmerken in de dataset staan van de gemeente en weer

andere kenmerken in de dataset van CBS. Er wordt als het ware een ‘saté-prikker’ door de verschillende datasets heen gehaald en de kenmerken worden gerelateerd aan dezelfde persoon (in veel gevallen gepseudonimiseerd of geanonimiseerd). De meest toegankelijke technologie dit dit in de praktijk mogelijk maakt heet Multi Party Computation. Deze technologie maakt gebruik van versleuteling van zowel de data als de algoritmes, waardoor er extra bescherming is tegen het (ongewild) openbaar maken van privacygevoelige informatie.

5.1 Federated Learning

Achtergrond

Federated Learning is een techniek die ook wel 'collaborative learning' wordt genoemd en vaak wordt toegepast met het doel om een algoritme voor kunstmatige intelligentie te trainen. Het uitgangspunt is het 'algoritme naar de data te brengen' (die dus bij de bron blijft). Een voorbeeld van Federated Learning is het trainen van een algoritme voor borstkanker-detectie. Dit algoritme analyseert beeld-data en andere patiëntgegevens die allemaal in één gezamenlijke dataset aanwezig zijn, bijvoorbeeld de data van een ziekenhuis. De analyse gebeurt binnen de context van het ziekenhuis en de data hoeft dus niet naar buiten verplaatst te worden. Vandaar ook de naam Federated Learning - het leren gebeurt lokaal. Dit proces gebeurt afzonderlijk op meerdere plaatsen. Na deze analyse worden de uitkomsten, zoals de input/gewichten van het algoritme, dus niet de data zelf, gecombineerd tot een gezamenlijke uitkomst. Dit is het Federated van Federated Learning.

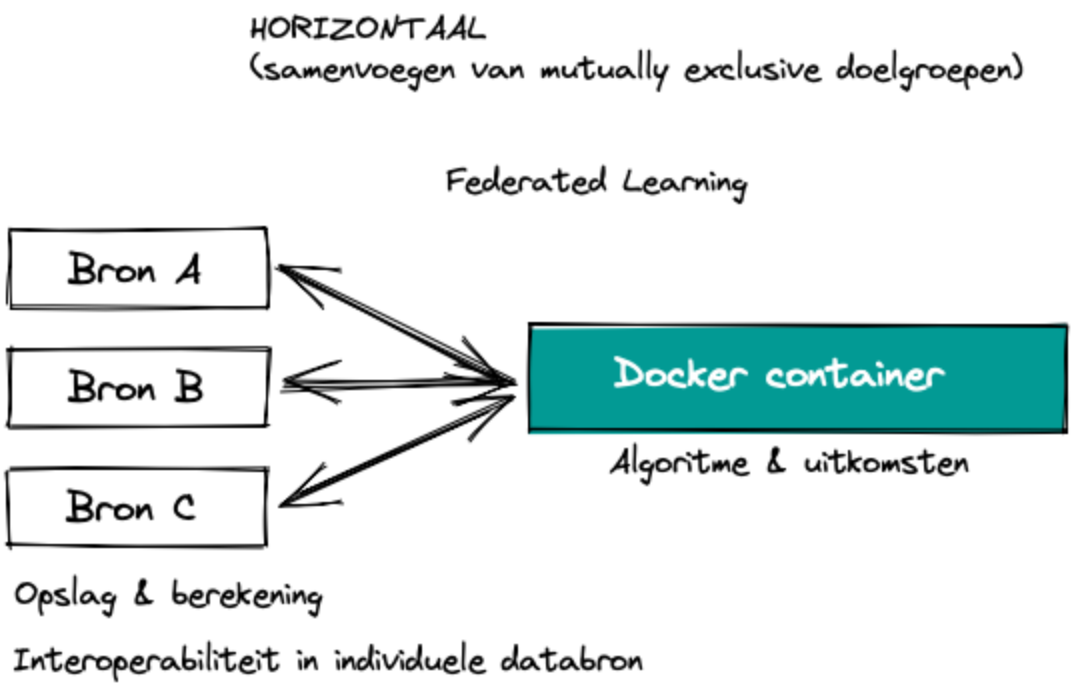
Federated Learning is één van de toepassingen van de Personal Health Train[8]. De Personal Health Train is een in Nederland door Health-RI ontwikkeld afsprakenstelsel dat het mogelijk maakt om vormen van gefedereerde data-analyse, zoals Federated Learning, te kunnen uitwisselen en op meerdere plaatsen toepasbaar te maken.

Federated Learning is een techniek die zeer goed toepasbaar is als alle kenmerken over personen of patiënten aanwezig zijn in een (al bestaande) dataset, zoals bijvoorbeeld het geval van patiënten in een ziekenhuis voor oncologisch onderzoek.

Technologie

Er zijn verschillende methoden voor Federated Learning. We behandelen niet alle mogelijkheden, maar richten ons specifiek op het in Nederland ontwikkelde afsprakenstelsel Personal Health Train (PHT). Dit afsprakenstelsel maakt het mogelijk om voor Federated Learning een gestandaardiseerde werkwijze te gebruiken. De verschillende stappen in een analyse worden bij de Personal Health Train uitgevoerd in een docker container, wat de mogelijkheid biedt om verschillende software(stappen) als één geheel te combineren. Zoals beschreven in paragraaf 4.1 maakt de docker container het makkelijker de analyse te verplaatsen van station naar station. Per station vindt de gewenste analyse plaats, die zich meestal richt op één database van één organisatie (of meerdere databases binnen de data governance structuur van één organisatie). De resultaten van de analyse worden vervolgens verstuurd naar een volgend station, bijvoorbeeld als input voor de 'gewichten' van een algoritme. Zie Figuur 13 voor een abstracte weergave.

[8] Zie voor meer informatie https://pht.health-ri.nl



Figuur 13. Deze weergave laat de relatie zien tussen een drietal bronnen (bron A, B en C) en Federated Learning. Er gaat geen data van de bron, maar er wordt een algoritme voor een specifieke bron gebruikt. De training van het algoritme gebeurt lokaal op de afzonderlijke bronnen. De uitkomsten, dus niet de data, worden gecombineerd tot een gezamenlijke uitkomst.

Het gebruik van docker containers als technologie lijkt vooral passend voor vraagstukken waar een uitgebreide lokale analyse plaatsvindt. Als voor iedere (simpele) analyse een docker container moet worden 'verplaatst' rijst de vraag of de overhead/impact van deze technologische keuze proportioneel is.

Data governance

Bij het toepassen van Federated Learning 'bezoekt' het algoritme (of in simpele gevallen de datavraag) de dataset van een betrokken partij. De gehele analyse vindt plaats binnen de data-omgeving van die partij. In de praktijk kun je dus zeggen dat de partij waar de analyse plaatsvindt optreedt als een Trusted Third Party'[9]. De partij is vanuit eigen data governance verplichtingen verantwoordelijk om de data op een juiste manier te beheren. Omdat data bij Federated Learning bij de eigenaar blijft betekent dit dat er voor de analyse geen nieuwe data governance hoeft te worden ingericht.

[9] Zie ook https://www.surf.nl/en/federated-learning

Beveiliging en juridische/privacy aspecten
Omdat data bij de bron blijft en alleen de analyses verplaatsen is Federated Learning inherent een zeer veilige technologie. Recent is er echter wel discussie over in hoeverre het mogelijk is om uitkomsten van Federated Learning analyses toch terug te voeren op personen[10].

De reikwijdte van de mogelijkheden van Federated Learning is afhankelijk van bestaande toestemming voor dataverwerking binnen organisaties. Er hoeven geen aparte, nieuwe verwerkersovereenkomsten te worden afgesloten, omdat individuele gegevens niet elders worden verwerkt. Er moet wel een (juridisch) kader worden afgesproken over wie de resultaten (op geaggregeerd niveau) teruggekoppeld krijgt en hoe deze verder verwerkt worden. Dit kan ook gebeuren bij één van de deelnemers (stations) binnen de Personal Health Train.

Ethische en sociale aspecten
Ethisch is Federated Learning een toepassing die sterk lijkt op onderzoek op eigen populaties en het delen van de (geaggregeerde) resultaten, met als doel beter inzicht te krijgen in diverse uitkomstmaten voor een populatie met dezelfde kenmerken die zich verspreid over meerdere locaties bevindt. Wat Federated Learning daarin toevoegt is een technische oplossing.

Relevantie geïntegreerde regionale data-infrastructuur
Minder relevant.

Data-architectuur
Voor regionale data-infrastructuur zijn gefedereerde analyses op afzonderlijke datasets minder interessant. Er worden voornamelijk toepassingen voorzien waarbij juist de combinatie van datavelden van dezelfde personen over verschillende datasets heen tot nieuwe inzichten leidt. Het regionale karakter leidt ertoe dat informatie van burgers uit verschillende zorgdisciplines samengebracht moet worden (verticale integratie van data), in tegenstelling tot het samenvoegen van elkaar uitsluitende populaties in dezelfde zorgdiscipline (horizontale integratie van data), denk aan ziekenhuisdata voor hetzelfde type aandoening in verschillende huizen.

Aggregatieniveau
Federated Learning maakt het weliswaar mogelijk om op individueel niveau te analyseren, maar niet zozeer over verschillende datasets heen.

Data-integratie
Vanuit data governance biedt Federated Learning een interessante aanpak, omdat het privacy-technisch, maar ook ethisch een methode is die inherent een aantal waarborgen biedt. Doordat het niet de methode is die focust op (verticale) integratie van data uit meerdere datasets, is Federated Learning minder toepasbaar in de regionale context.

[10] https://arxiv.org/abs/2112.02918

5.2 Multi Party Computation

Achtergrond

Wanneer we data willen analyseren over *dezelfde persoon* maar uit *verschillende datasets,* hebben we een andere aanpak nodig dan het hiervoor beschreven Federated Learning. We moeten in dit geval data tussen verschillende datasets laten 'reizen', maar voorkomen dat er privacy-gevoelige informatie gelekt wordt. De oplossing hiervoor ligt in het versleutelen van de data uit iedere bron voordat de data tussen de verschillende bronnen gedeeld wordt. De technologie die hiervoor beschikbaar is wordt Multi Party Computation (MPC) genoemd. MPC is ontstaan eind jaren '80 en is flink doorontwikkeld in de afgelopen 10 jaar. Onder andere TNO heeft in Nederland afgelopen jaren geïnvesteerd in de ontwikkeling[11] en er zijn diverse succesvolle pilots gedaan (met o.a. CBS, CZ en Zuyderland) binnen de zorg[12].

TNO beschrijft het als volgt in de whitepaper over Privacy Preserving Analytics[13]: *"Door gebruik te maken van geavanceerde cryptografische technieken kan MPC gevoelige data nog beter beschermen dan FL. MPC-protocollen zorgen er namelijk voor dat alle data in versleutelde vorm blijft. De berekeningen worden uitgevoerd op de versleutelde data en alleen de uitkomst van de analyse wordt ontsleuteld. Dus zelfs terwijl de data te allen tijde versleuteld blijft kan er mee gerekend worden. Hiermee realiseert MPC een maximaal haalbare mate van privacy en vertrouwelijkheid; alleen de uitkomst van de analyse wordt onthuld. Een nadeel is dat MPC over het algemeen meer rekenkracht en/of een zwaardere communicatie infrastructuur nodig heeft dan FL."*

Multi Party Computation - een simpel voorbeeld

Stel: er zijn vier collega's die allemaal hetzelfde werk doen en die willen uitrekenen wat ze gemiddeld verdienen, zodat ze weten of ze zelf boven of onder het gemiddelde verdienen. Maar geen van hen wil de rest vertellen hoeveel hij of zij precies verdient. Hoe berekenen ze dan het gemiddelde? Dit kan door 'secret sharing' te gebruiken (een onderdeel van MPC): ieder salaris wordt verdeeld in vier 'stukjes', zodanig dat de optelsom van de vier stukjes gelijk is aan het totaal salaris van die persoon. Vervolgens worden die stukjes willekeurig verdeeld over de collega's, zodat iedereen vier willekeurige stukjes heeft. Vervolgens telt iedere persoon de willekeurige stukjes die hij heeft gekregen bij elkaar op. Het totaal zegt niets meer over het salaris van die persoon. Maar door al die willekeurige totalen bij elkaar op te tellen en te delen door 4 kan wel het gemiddelde salaris worden berekend. Zo kunnen de collega's hun gemiddelde salaris berekenen en weten of zij onder of boven het gemiddelde verdienen, zonder dat ze elkaar hebben hoeven te vertellen hoeveel ze precies verdienen.

[11] https://www.tno.nl/nl/aandachtsgebieden/informatie-communicatie-technologie/roadmaps/data-sharing/multi-party-computation/
[12] https://privacy-web.nl/wp-content/uploads/po_assets/560647.pdf
[13] https://www.tno.nl/nl/aandachtsgebieden/informatie-communicatie-technologie/roadmaps/data-sharing/secure-multi-party-computation/

Technologie

Om MPC te implementeren is er (nog) geen eenduidige overkoepelende standaard. Hier wordt wel aan gewerkt door standaardisatie-organisaties[14]. Praktisch gezien zijn MPC protocollen echter wel veelal openbaar en vrij toegankelijk in de academische literatuur. Voorbeelden hiervan zijn het MPyC[15] pakket (bestaande uit diverse MPC protocollen) en diverse protocollen voor 'secure comparison' operaties[16].

MPC implementaties van dergelijke protocollen zijn beschikbaar via open source software. In de praktijk worden deze implementaties nu nog aangeboden door een beperkt aantal aanbieders, waaronder in Nederland Linksight (TNO spinoff), Roseman Labs (TU/e spinoff) en in het buitenland o.a. Partisia (Denemarken).

Een implementatie van MPC bestaat uit verschillende onderdelen, als voornaamste een MPC netwerk en meerdere MPC nodes. Deze begrippen worden hieronder kort toegelicht:

- **MPC netwerk.** Organisaties die data willen analyseren vormen een gezamenlijk netwerk. Binnen dit netwerk worden de (versleutelde) gegevens die nodig zijn voor analyses gedeeld. Binnen implementaties van MPC wordt dit ook wel een 'compute group' of een 'privacy engine' genoemd.
- **MPC node**. Een MPC netwerk bestaat uit nodes, vaak één per deelnemende organisatie. Deze nodes vormen de connectie van de organisatie met het MPC netwerk voor het aanleveren van de input van de deelnemende partijen.

Alle partijen die data aanleveren binnen het MPC netwerk doen lokaal de dataversleuteling. Voor het uitvoeren van berekeningen op deze versleutelde gegevens binnen een MPC netwerk worden speciale algoritmes gebruikt. Deze algoritmes zijn in staat met versleutelde gegevens te rekenen in plaats van met 'normale' gegevens. Dat betekent wel dat voor MPC berekeningen vaak meer rekenkracht nodig is. MPC kan op verschillende manieren geïmplementeerd worden. Afhankelijke van de aanbieder zijn er implementaties die het bijeenbrengen van data bij een derde partij doen (buiten het analyse-netwerk) of juist binnen het eigen analyse-netwerk. De uitkomsten van de MPC berekening worden door middel van datavirtualisatie samengebracht tot één resultaat dat aan de deelnemende partijen wordt teruggekoppeld, zonder dat de onderliggende gegevens daarvoor hoeven te worden gedeeld (in het voorbeeld: het gemiddelde salaris wordt teruggekoppeld, de gebruikte gegevens niet). Zie voor een schets Figuur 14.

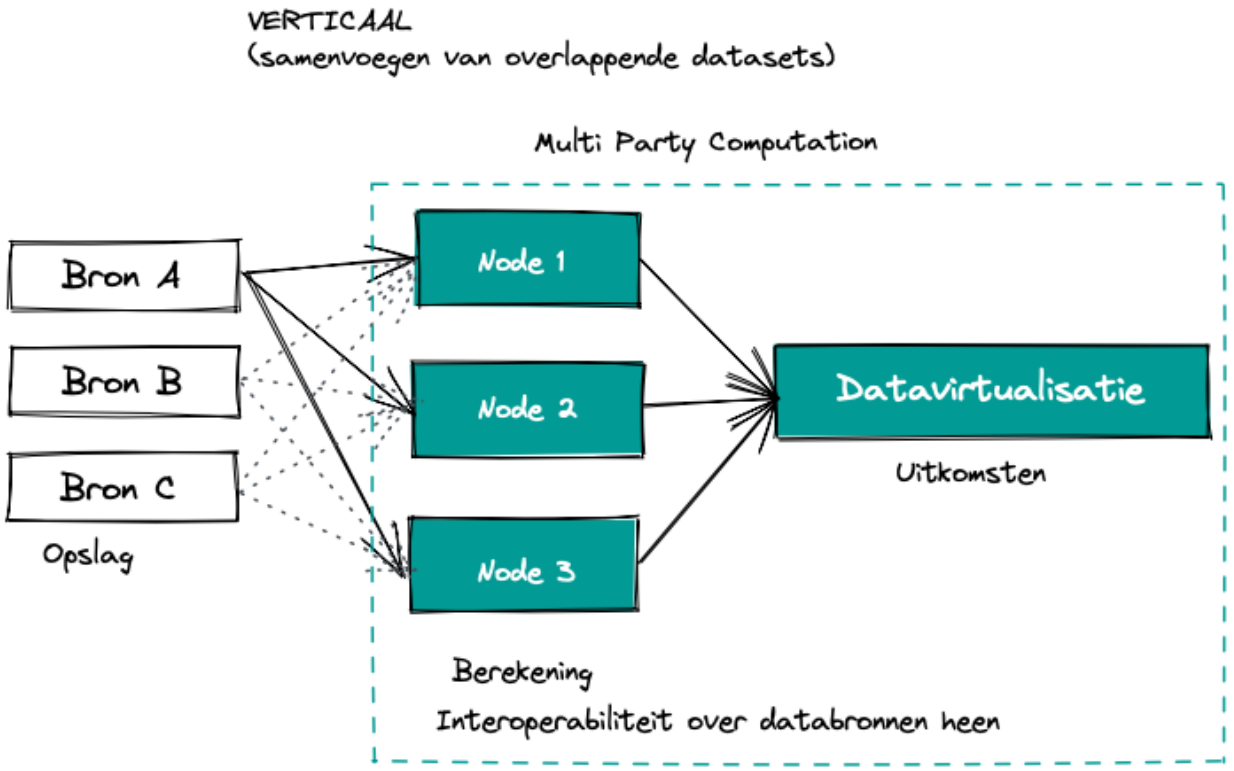
---

[14] NIST: https://csrc.nist.gov/projects/threshold-cryptography, IETF: https://www.ietf.org/archive/id/draft-patton-cfrg-vdaf-01.html, ISO: https://www.iso.org/standard/80508.html

[15] https://www.win.tue.nl/~berry/mpyc/

[16] - Damgard et al. (DFK+06): Section 5 presents the "Less than" operator, and section 7 the equality test: https://iacr.org/archive/tcc2006/38760286/38760286.pdf
- Schoenmakers et al. (GSV07): Secure comparison of integers: https://www.win.tue.nl/~berry/papers/pkc07intcomp.pdf

Omdat iedere deelnemende organisatie binnen een MPC netwerk een MPC node moet implementeren vraagt dit actieve betrokkenheid van iedere partner. Dit kan gaan om het (handmatig) aanleveren van data via een simpele upload module of het inregelen van een 'live' interface. Er zijn ook aanbieders die MPC aanbieden als software-as-a-service.



Figuur 14. Deze weergave laat de relatie zien tussen een drietal bronnen (Bron A, B en C) die gekoppeld zijn aan een drietal servers (Server 1, 2 en 3) en vervolgens samenkomen in een virtuele laag (in dit geval datavirtualisatie). De groene gestippelde lijn laat zien dat een combinatie van deze servers en de datavirtualisatie gebruik maakt van Multi Party Computation. Ook bij PPA wordt er geen data opgeslagen: de data blijft bij de bron. Een gebruiker kan via de virtuele laag onderzoekshandelingen uitvoeren.

Data governance

De data governance bij MPC is deels technisch en deels juridisch. Technische oplossingen (zoals een audit netwerk/logging) maken het mogelijk om kaders te stellen aan de vraag die wordt gesteld en de uitkomsten die worden gedeeld. Dit maakt het transparant welke data op welke wijze wordt gebruikt. Daarnaast zorgt het gebruik van uitsluitend versleutelde gegevens dat (herleidbare) persoonsgegeven in het kader van AVG anders geïnterpreteerd kunnen worden en er daardoor een andere (mogelijk lichtere) governance nodig is. In verband met de schaal van de data-verwerking zal echter vrijwel in alle gevallen alsnog een data protection impact assessment (DPIA) nodig zijn, evenals afspraken over aanlevering, beheer en gebruik van de data. Dat maakt dat dus bij MPC de invulling wellicht anders kan dan bij bijvoorbeeld een regionaal datalake, maar dat nog altijd duidelijke (juridische) afspraken nodig zijn met betrekking tot de data governance.

Beveiliging en juridische/privacy aspecten

In de basis is MPC een techniek die uitgaat van privacy by design. Tot welk niveau dit is doorgevoerd kan afhangen van de specifieke implementatie. Een DPIA is altijd nodig als MPC op grotere schaal wordt toegepast (tussen meerdere organisaties).

Het 'voordeel' van MPC met betrekking tot privacy is dat het een technische oplossing (bijvoorbeeld via een audit-netwerk) biedt voor diverse waarborgen die de AVG juridisch probeert af te dichten. Zo worden binnen MPC uitsluitend versleutelde gegevens gebruikt, die niet tot personen herleidbaar zijn. Daarnaast kan binnen MPC, afhankelijk van de gekozen implementatie, een technische oplossing worden ingericht waarin governance-afspraken over bijvoorbeeld de onthulling criteria worden vastgelegd. Dit in tegenstelling tot bijvoorbeeld een regionaal of nationaal datalake, waarbij enkel uitgegaan wordt van juridische afspraken.

Omdat gegevens alleen gebruikt worden voor de duur van de analyse en alleen versleuteld (kunnen) worden opgeslagen zijn er andere beveiligingseisen dan aan een omgeving en/of database waarin onversleutelde gegevens worden opgeslagen.

Ethische en sociale aspecten

MPC biedt de meest uitgebreide privacy by design mogelijkheden van alle toepassingen die worden beschreven in deze whitepaper. Het is echter ook een relatief nieuwe technologie vergeleken met de andere mogelijkheden. Een goede implementatie van MPC vraagt dan ook om zorgvuldige aandacht van betrokkenen om de juiste keuzes te maken en alle deelnemende partijen en medewerkers mee te nemen in de manier waarop MPC voor een veilige oplossing kan zorgen. De veiligheid en het 'gemak' waarmee MPC het mogelijk maakt om persoonlijke gegevens te analyseren is namelijk tegelijkertijd ook een risico vanuit ethisch opzicht. Als het *te* gemakkelijk wordt om gegevens bij elkaar te brengen terwijl er geen duidelijk ethische en maatschappelijke kaders zijn voor de context waarbinnen een onderzoek plaats mag vinden, ontstaat het risico op ongewenst en onethisch gebruik van data binnen het netwerk. Juist bij een technologie waar (theoretisch) veel mogelijk is, is het extra belangrijk om als deelnemende partijen heldere kaders mee te geven.

Relevantie geïntegreerde regionale data-infrastructuur

MPC lijkt de meest vergevorderde oplossing om op een privacy-vriendelijke manier gegevens uit meerdere bronnen te analyseren, zonder een centrale opslag van gegevens te moeten creëren. In vergelijking met bijvoorbeeld een nationaal data lake zijn er meer mogelijkheden om laagdrempelig organisaties aan te sluiten en binnen de regio zelf de kaders te creëren voor onderzoek. Tegelijkertijd is dat ook de grootste uitdaging bij het gebruik van MPC. Een regio moet zelf duidelijke werkwijzen en afspraken maken en kunnen niet 'zomaar' leunen op de infrastructuur van bijvoorbeeld bestaande partijen zoals CBS, die hiervoor een bestaande infrastructuur hebben (echter wel binnen andere, meer nationaal-gerichte kaders).

Data-architectuur
De decentrale opzet van MPC maakt het mogelijk een privacy als centraal uitgangspunt te hanteren en data zoveel mogelijk bij de bron te laten. De inrichting van de data architectuur biedt daarnaast mogelijkheden om relevante, tijdige data te gebruiken en zo maximaal aan te sluiten op regionale doelstellingen.

Aggregatieniveau
Met behoud van privacy is het mogelijk op n=1 niveau data-analyses uit te voeren. Dit is echter wel ingewikkelder ('zwaarder') dan normale analyses met bijv. niet-versleutelde, centrale data.

Data-integratie
MPC is een oplossing die zeer goed toepasbaar is voor het integreren van data over meerdere datasets van verschillende partijen door het gebruik van gekoppelde nodes. Voorbeeld-projecten hebben uitgewezen dat dit ook goed toepasbaar kan zijn in de zorg.

6 Overzicht vergelijking

In Tabel 1 zijn de vijf mogelijkheden voor geïntegreerde data-infrastructuur nogmaals op een rij gezet met de negen aspecten waarop getoetst is. Deze tabel dient zowel als samenvatting van de belangrijkste punten, maar geeft ook weer op welke aspecten de mogelijkheden overeenkomen en verschillen.

Tabel 1: Deze tabel is een samenvatting van de vijf mogelijkheden voor geïntegreerde data-infrastructuur en de negen aspecten waarop getoetst is. Er is geprobeerd een zo volledig mogelijke samenvatting per mogelijkheid te geven om de mogelijkheden te vergelijken.

6.1 Achtergrond

	Nationaal datalake	Regionaal datalake	FAIR data station	Data-virtualisatie	Federated Learning	Multi Party Computation
Doel	Statistiek en onderzoek	Onderzoek en datagedreven beleid	Valide data vragen aan de bron aanbieden ter verwerking.	Tijdelijk samenbrengen van data in een virtuele laag.	Analyse van data op afzonderlijke datasets als input voor machine learning (bijv. AI)	Privacy behouden tijdens analyses binnen eigen analyse-omgeving
Volwassen-heid	Bewezen technologie, ook voor gezondheids-onderzoek.	Bekende technologie, maar nieuw op te zetten regionale infrastructuur en data governance.	Toepassing bij KIK-V en Federated Learning. Een infrastructureel netwerk van data stations bestaat al langer, gefedereerd berekeningen uitvoeren is relatief nieuwe technologie.	Steeds meer ingezet in de zorg, ook bij RUG-CIT. Soms ook toepasbaar als deel van een oplossing	Bewezen concept. Vooral toegepast binnen medisch domein voor AI en big data-onderzoek (bijv. imaging)	Relatief nieuw. Afgelopen 5 jaar meerdere implementaties nationaal en internationaal
Schaal-baarheid	Oplossing minder schaalbaar (projectmatige structuur en complexiteit beveiliging grote hoeveelheden (zorg)data).	Oplossing minder schaalbaar (projectmatige structuur en complexiteit beveiliging grote hoeveelheden (zorg)data).	Schaalbaarheid wordt geborgd door het federatieve karakter, dat voorkomt dat veel data centraal wordt opgeslagen. Ook kan er sneller	Oplossing is technisch schaalbaar. Afhankelijk van reikwijdte van de toepassing is gevraagde infrastructurele capaciteit groot.		Afgelopen jaren betere mogelijkheden encryptie/ schaalbaarheid toegenomen

			opgeschaald worden doordat de oplossing technisch bestaat uit een node per dataleverancier waardoor eenvoudig gefaseerd opgeleverd kan worden.			

6.2 Technologie

	Nationaal datalake	Regionaal datalake	FAIR data station	Data-virtualisatie	Federated Learning	Multi Party Computation
Opslag	Centrale opslag.	Centrale opslag.	Geen centrale opslag. Data blijft bij de bron.	Geen centrale opslag. Data blijft bij de bron.	Geen centrale opslag. Data blijft bij de bron.	Geen centrale opslag. Data blijft bij de bron.
Technologie	Statische bronbestanden + tooling beschikbaar via remote access omgeving.	(Cloud-based) database technologie en BI platform.	3 toepassingen die variëren in complexiteit en beveiliging.	Beveiligde koppelingen en virtuele omgeving voor gebruikers.	Container technologie (Docker).	Multi Party Computation (MPC).
Aggregatie-niveau	Aggregatieniveau output n>10.	Aggregatieniveau output n>10.	Aggregratie-niveau n=>1.	Aggregratie-niveau n=1.	Aggregatieniveau naar eigen keuze/afspraken.	Aggregratieniveau n=1
Niveau technische kennis	Geen specifieke technische kennis/ implementatie nodig per deelnemende partij.	Enige technische kennis/ implementatie nodig per deelnemende partij (voor inrichten omgeving)	Technische implementatie is geconcentreerd rondom de installatie van de data stations en programmeertaal van het gekozen type data station	Bestaande technologie die aan te schaffen en te configureren is	Focus op 'zwaardere' toepassing zoals machine learning	Relatief eenvoudige technische kennis/ implementatie nodig per deelnemende partij.

6.3 Data governance

	Nationaal datalake	Regionaal datalake	FAIR data station	Data-virtualisatie	Federated Learning	Multi Party Computation
Voorzieningen	Nationale voorziening	Zelf in te richten, regionale samenwerking	Af te stemmen per data stations in het netwerk.	Zelf in te richten o.b.v. need to know.	Zelf in te richten o.b.v. need to know.	Zelf inrichten, combinatie van technische en juridische afspraken
Structuur	Projectmatige structuur	Projectmatige structuur binnen eigen regio	In te richten in staande organisatie	In te richten in staande organisatie.	Projectmatige structuur	Projectmatige structuur binnen eigen regio
Beheer	Centrale governance op gestructureerde wijze door (zeer) uitgebreide CBS beheerorganisatie.	Centrale governance met data-analist in dienst van de stichting	Beheer door betreffende organisaties.	Te bepalen waar virtuele oplossing beheerd wordt.	Meestal beheer door uitvoerder onderzoek of beheersorganisatie kennispartij	Gezamenlijk door deelnemende organisaties, trusted computing environment bij deelnemende of externe partij
Onderzoeks-protocol	Verplicht onderzoeks-protocol aan te leveren (met goedkeuring METC).	Ja, gezamenlijk opgesteld onderzoeks-protocol	Volgt afsprakenset (KIK-V) en technische inrichting van attest	Regionaal in te richten	Gezamenlijk opgesteld	Gezamenlijk opgesteld

6.4 Beveiliging en juridische/privacy aspecten

	Nationaal datalake	Regionaal datalake	FAIR data station	Data-virtualisatie	Federated Learning	Multi Party Computation
Data bescherming	Pseudonomisatie van bestanden (BSN -> RIN) Data wordt via een beveiligde verbinding verzonden.	Pseudonomisatie (dubbel) van bestanden (via TTP) Data wordt via een beveiligde verbinding verzonden.	Databerekening vindt in de bron plaats. Alleen in meest complexe variant is beveiliging van data en algoritme een optie.	Primair geen anonimisatie of encryptie. Biedt een veilige wijze van samenvoegen van data.	Analyses vinden plaats bij de bron, data wordt niet verplaatst, alleen uitkomsten	Toegang van versleutelde datapunten via specifieke nodes. Afhankelijk van implementatie trusted computing environment in netwerk of extern
Toegang	Persoons-gebonden toegang	Persoons-gebonden toegang	Volgt de afsprakenset (KIK-V) en de afspraken tussen organisaties en de gemandateerde datavragers	Zelf in te richten, persoons-gebonden toegang en bewerking	Toegang via beveiligde koppelingen	Op basis van technische afspraken MPC/audit nettwerk
Logging	Logging en maatregelen bij misbruik door CBS.	Logging en maatregelen bij gebruik.	Ingebed in de oplossing.	In te richten in virtuele omgeving.	Zelf in te richten op basis van implementatie	Audit-netwerk (bijv. op basis van blockchain)
Output-controle	Outputcontrole door CBS	Zelf in te regelen	Outputcontrole: uitkomsten worden gedeeld. Accent ligt echter bij de inputcontrole van de gevraagde onderzoeksresultaten (in het 'attest')	Zelf outputcontrole af te spreken in governance en technisch te faciliteren waar nodig.	Geen standaard voorziening voor output-controle	Outputcontrole vooraf vastgesteld. (Eisen aan uitkomsten van data kunnen vooraf worden vastgesteld.)
Overeenkomst	Juridische overeenkomst voor gebruik omgeving	Gezamenlijke juridische entiteit deelnemers (stichting). Verwerkersover-eenkomst tussen betrokkenen.	Notariseren van queries, personen en resultaten.		Op basis van onderzoeksopzet	Notariseren van queries en resultaten door per partij bijv. Criteria te definieren.

Informed consent	Geen informed consent nodig (valt onder de Wet op Statistiek).	Grotendeels passief informed consent (obv onderzoeks-gronden)	Grotendeels passief informed consent (obv onderzoeks-gronden)	Informed consent zelf in te richten op basis van geldende regelgeving.	Informed consent.	Informed consent. Meer 'dynamische' onderzoeken mogelijk, maar wel vraag hoe je dit proces goed inricht .

6.5 Ethische en sociale aspecten

	Nationaal datalake	Regionaal datalake	FAIR data station	Data-virtualisatie	Federated Learning	Multi Party Computation
Borging procedures	Onderzoeks-voorstel naar METC.	Onderzoeks-voorstel	In de afsprakenset.	Onderzoeksproto collen regionaal opgesteld, ingevoerd en gemonitord.	Afhankelijk van de onderzoeksopzet.	Grotendeels technisch af te dwingen als onderdeel van aanpak MPC
Zeggenschap/ governance	Weinig zeggenschap. Vast staande procedures.	Veel zeggenschap. Zelf in te richten procedures.	Organisatie beoordeelt elk verzoek, dus heeft maximale zeggenschap.	Zelf in te richten.	Zelf in te richten	Zelf in te richten in de regio; aansluitend op regionale doelstellingen

6.6 Relevantie geïntegreerde regionale data-infrastructuur

	Nationaal datalake	Regionaal datalake	FAIR data station	Data-virtualisatie	Federated Learning	Multi Party Computation
Relevantie	Minder relevant.	Minder relevant.	Relevant maar beperkt toepasbaar.	Zeer relevant; voor zorgdata in combinatie met FL of MPC.	Minder relevant.	Zeer relevant.
Data-architectuur	Technisch gezien een geschikte en relevante optie als opslagvorm. Hiervoor is het echter wel nodig	Technisch gezien een geschikte en relevante optie als opslagvorm. Hiervoor is het echter wel nodig	Door middel van de verschillende data stations kunnen datadragers en dataleveranciers	Biedt technisch de mogelijkheden om gefedereerde data te gebruiken. Belangrijk aspect	Focus op machine-learnin g en zwaardere toepassingen binnen de context van	Privacy als centraal uitgangspunt, data zoveel mogelijk bij de bron, goede mogelijkheden

	dat veel data verplaatst wordt.	dat veel data verplaatst wordt.	informatie uitwisselen. De complexiteit van de techniek hangt af van het benodigde niveau van beveiliging van data en algoritmen	daarbij is dat er centraal geen data wordt opgeslagen.	verschillende organisaties met hetzelfde type personen (horizontale integratie)	gebruik van tijdige data.
Data-Aggregatie-niveau	CBS-richtlijnen beperken output op aggregatie-niveau lager dan n=10.	Technisch gezien geschikt voor aggregatie-niveau n=1. Praktisch gezien is detailniveau van de brondata bepalend. Transmuraal Netwerk kiest voor beperken output op aggregatie-niveau lager dan n=10.	Sparql en RESTful acteren op het niveau van geaggregeerde data. Het docker container voorziet in datavragen en -antwoorden op individueel niveau	Data uit de bronnen wordt op individueel niveau geaggregeerd op basis van unieke identifiers.	Mogelijk om op individueel niveau te analyseren, maar alleen binnen de context van één organisatie	Mogelijkheid voor analyse op n=1 niveau met behoud van privacy; tevens ook mogelijk om andere relevante data waar nodig te gebruiken; Output afhankelijk wettelijke kaders en regionale afspraken
Data-integratie	Technisch gezien is het mogelijk om zowel horizontaal als verticaal onderzoek toe te passen.	Technisch gezien is het mogelijk om zowel horizontaal als verticaal onderzoek toe te passen.	Technisch gezien is het mogelijk om zowel horizontaal als verticaal onderzoek toe te passen (voor de variant Docker Container).	Datavirtualisatie voorziet in horizontale en verticale data-integratie. Het is een maatwerk-oplossing die technisch en qua onderzoeksprotocollen zelf gecreëerd en beheerd moet worden.	Gericht op integratie van uitkomsten van data-analyses uit meerdere datasets met dezelfde type informatie over bijv. personen	Goede mogelijkheden voor integratie van data over meerdere datasets

7. Conclusie

Om te komen tot de beste geïntegreerde regionale data-infrastructuur voor de Achterhoek doorlopen we, op basis van de in paragraaf 1.2 beschreven specificaties voor GERDA, het door ons opgestelde schema met de drie ontwerpkeuzes (en bijbehorende toepassingen) zoals volledig weergegeven in Figuur 15 hieronder.

Ontwerpkeuze 1: *data-architectuur*
Op basis van de uitgangspunten 'data blijft zoveel mogelijk bij de bron' (DIZRA principe) en 'zo min mogelijk handmatige handelingen' (uitgangspunten efficiënt en veilig) wordt gekozen voor een **gefedereerde data-architectuur**. De auteurs schatten in dat gezien de gewenste hoeveelheid te integreren data een centrale data-architectuur te risicovol en daarmee te beperkt schaalbaar is.

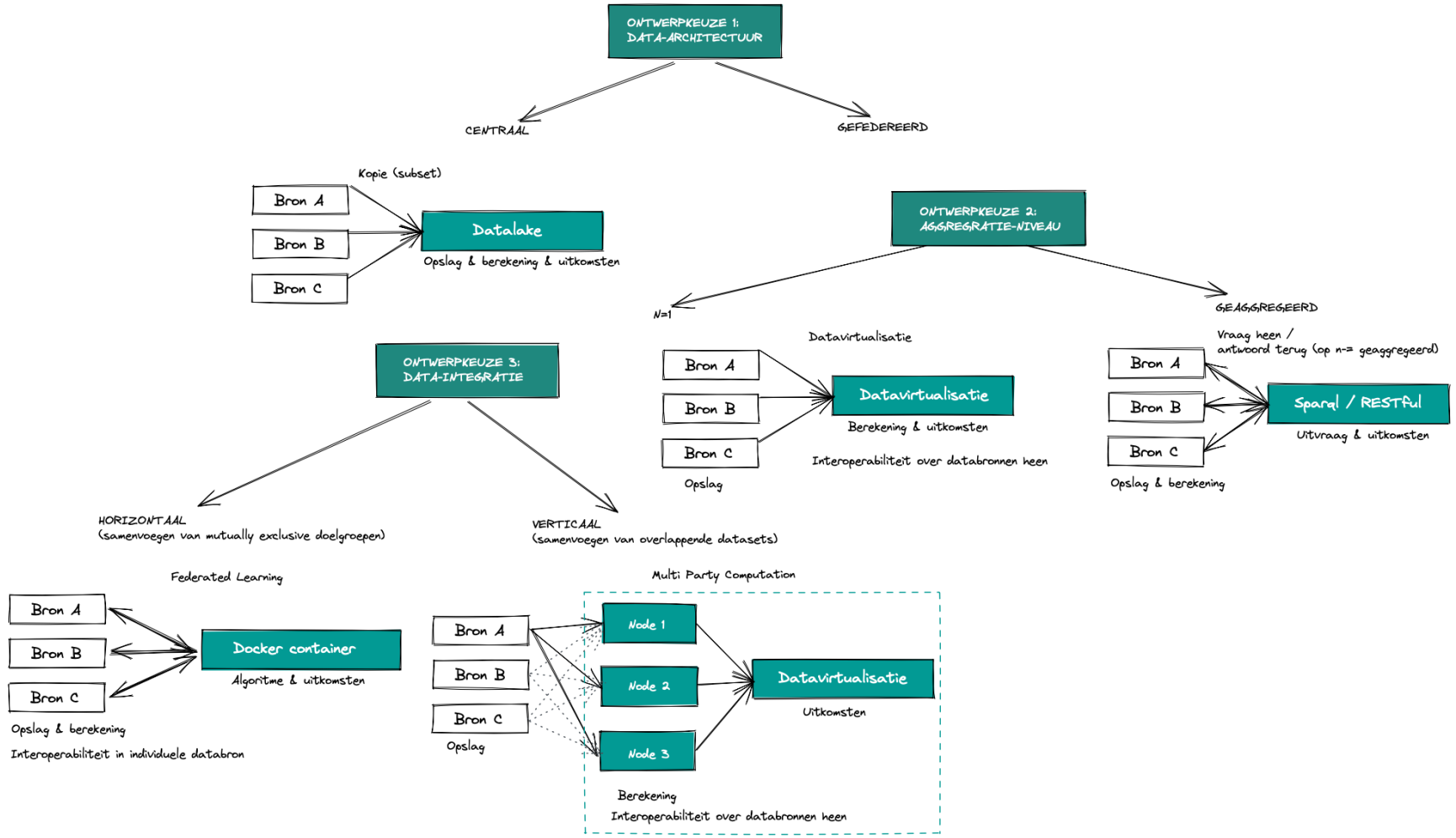
Ontwerpkeuze 2: *aggregatieniveau*
Omdat de data 'faciliterend moet zijn voor gebruik op persoonsniveau' vanwege 'wederkerig gebruik' voor primair en secundair gebruik (uitgangspunten relevant voor de regio) wordt gekozen voor een gewenst **aggregatievniveau van n=1**.

Ontwerpkeuze 3: *data-integratie*
Tot slot wordt, omdat het gaat om analyse van zorgdata (en daarmee bijzondere persoonsgegevens) op n=1 niveau, gekozen voor een techniek die uitgaat van privacy-by-design (uitgangspunt efficiënt en veilig) en die geschikt is voor **verticaal onderzoek**; het samenvoegen van verschillende data van één persoon over bronnen heen.

Multi Party Computation (in combinatie met datavirtualisatie) is daarmee de meest relevante techniek om verder te onderzoeken en mee te nemen in een technisch ontwerp voor GERDA. Daarbij aangetekend dat binnen de overige oplossingen (o.a. FAIR data stations) elementen zitten die op onderdelen ook kunnen bijdragen aan de gewenste regionale oplossing.

In het volgende werkpakket van de werkgroep GERDA i.s.m. Population Health Data NL wordt nader uitgewerkt welke gebruikersgroepen - met welke tooling - in de data-werkplaats tot een antwoord kunnen komen. Daarnaast zal onderzocht worden of het mogelijk is een *proof of concept* van MPC in te richten rondom acute zorg in de Achterhoek.

Wil je bijdragen aan de (door)ontwikkeling van het technische ontwerp van GERDA, laat het ons dan weten. Neem hiervoor contact op met Ilona Oude Nijhuis (ilonaoudenijhuis@room-to.nl) of Maarten den Braber (maarten@populationhealthdata.nl)



Figuur 15. Overzicht van de drie ontwerpkeuzes met de bijbehorende mogelijkheden zoals deze zijn uitgelegd in deze whitepaper.